Some Problems in Curve and Surface Estimation

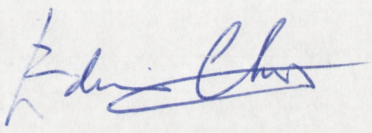
Chik Wan Edwin Choi

A thesis submitted for the degree of Doctor of Philosophy of
The Australian National University

April 1998

Declaration

Unless otherwise specified in the text, this thesis described my own work, supervised by Professor P.G. Hall and published jointly with him.



Chik Wan Edwin Choi

Acknowledgements

I would like to express my deep felt gratitude to my supervisor, Professor Peter Hall for his constant guidance and support. His insightful comments and advice have led me through some of the toughest times in my research. I am also indebted to Peter for his assistance in my scholarship applications, and for giving me the opportunity to visit the Victoria University at Wellington, where parts of this thesis were written. I could not have a better supervisor than him.

I would also like to thank Dr. Daniel Lunn for introducing statistics to me when I was an undergraduate, and for recommending me to be Peter's research student.

I am grateful to the following people:

– Dr. Berwin Turlach reviewed parts of this thesis, and provided constructive criticism on the presentation.

– Steve Davies helped me convert my TEX files to LATEX files, and taught me how to do most of the layout.

– Professor David Vere-Jones and Dr. David Harte provided the earthquake data and some of the `Splus` functions used in Chapters 5 and 6. I would like to thank them for their hospitality while I was visiting the Victoria University at Wellington, where I benefitted enormously from the fruitful discussions with them about earthquakes.

I acknowledge the financial support from the School of Mathematical Sciences and the Australian Government Department of Employment, Education and Training in funding an Australian National University PhD Scholarship and an Overseas Postgraduate Research Scholarship respectively. Lastly, I would like to thank my parents for the moral support throughout the years.

Abstract

The main theme of this thesis is nonparametric curve and surface estimation. The first four chapters concentrate on the former problem, where a new technique is introduced which improves on the bias of conventional local linear smoothers in regression analysis and two-parameter locally-parametric estimators in density estimation. Our method involves calculating an estimate of the regression function or density at a point which is close to the point x at which we wish to estimate the curve, and using this estimate to evaluate an approximation at x. A list of estimators exploiting this methodology is proposed, and may be shown to reduce bias by up to two orders of magnitude. Finite-sample properties of our new estimators are investigated in simulation studies.

The last two chapters focus on nonparametric surface estimation, where the surface represents the intensity of a point process in the plane. The surface contains poles, which correspond to places where the intensity is asymptotic to infinity. Statistical methods for estimating the location and "strength" of a pole are developed. In particular, it is shown how the correlation dimension, a well-known quantity in the fractal context, of a point process in the neighbourhood of the pole is related to the "strength" of the pole. The techniques are illustrated with earthquake data taken from the Kanto region in Japan.

Related Publications

The following papers have been submitted for publication from the work in this thesis:

Choi, E. and Hall, P. (1997). On the estimation of poles in intensity functions. Submitted to *Biometrika*.

Cheng, M.Y., Choi, E., Fan, J. and Hall, P. (1998). Skewing-methods for two-parameter locally-parametric density estimation. Submitted to *Bernoulli*.

Choi, E. and Hall, P. (1998). On bias reduction in local linear smoothing. *Biometrika*. To appear.

Contents

Chapter 1

Kernel Regression

1.1 Introduction

Nonparametric regression provides a useful tool for studying relationships between covariates and responses in regression analysis. In nonparametric regression, we remove the restriction that the underlying curve of interest belongs to a pre-determined class of functions that depend on a finite number of parameters. This approach is particularly attractive when we have little prior knowledge about the structure of the data. Admittedly, nonparametric estimators have zero asymptotic efficiency compared to parametric estimators when the true model is employed. Nevertheless, fitting incorrect regression models leads to inconsistent curve estimators, even if we have plenty of data.

Basically, the form of regression is determined by the model in parametric regression, and is driven by the data in nonparametric regression. Because of this, a pre-specified parametric model is often too restrictive to be able to pick up unexpected features of the regression function. A nonparametric approach, on the other hand, provides a flexible method for exploring general relationships between variables. A landmark example is the study of human longitudinal height growth curves in which the first derivative of the regression function (which corresponds to the rate of height growth) is of interest (see for example, Gasser *et al.*, 1984; Ramsay and Silverman, 1997). The nonparametric method is able to pick up an extra peak in the first derivative which indicates a mid-growth spurt at the age of about seven. This peak is difficult to detect by *ad hoc* parametric models, unless one has incorporated this knowledge as part of the models. Although this example demon-

strates convincingly the merits of nonparametric regression, it should be noted that parametric and nonparametric methods are by no means mutually exclusive competitors. Quite often, it is possible to suggest simple parametric relationships from the nonparametric analysis. Moreover, in cases where we have information on the form of the underlying regression function, it proves to be useful to employ nonparametric regression techniques to consolidate or justify our prior understanding of the curves. See, for example, the monograph by Hart (1997).

There is now a variety of methods for obtaining nonparametric curve estimators, some of which are intuitively simple and some mathematically sophisticated. Current nonparametric techniques employed are mainly based on kernel functions, splines and wavelets. Recent introductions to kernel and spline approaches may be found in the monographs by Wand and Jones (1995) and Green and Silverman (1994), and on wavelets in the paper by Nason and Silverman (1997). Kernel methods are arguably the simplest in terms of interpretability among the three mentioned methods, and we shall review the most relevant ones in Sections 1.2 and 1.3. We shall compare the Nadaraya-Watson estimator, the Gasser-Müller estimator and the local polynomial estimator, in terms of their theoretical and practical performances. Section 1.4 will discuss some bias-reduction techniques for general kernel methods, and Section 1.5 will outline some contemporary devices for guarding against sparse design in local linear smoothing.

1.2 Kernel Estimators

Unless otherwise stated, we assume that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent and identically distributed random variables, with respective conditional regression mean and variance given by

$$m(x) = E(Y|X = x) \quad \text{and} \quad v(x) = \text{var}\,(Y|X = x)\,,$$

where (X, Y) denotes a generic pair of random variables from the sample. Let $f(x, y)$ be the joint density of X and Y, and $f_X(x)$ be the marginal density of X. Our aim is to estimate the regression function m. To measure the closeness of the true regression mean and its estimate $\widehat{m}$ locally at x, we use the mean squared error (MSE) criterion which is defined as $\text{MSE}\{\widehat{m}(x)\} = E\{m(x) - \widehat{m}(x)\}^2$. Note that $\text{MSE}\{\widehat{m}(x)\}$ can be decomposed as the sum of squared bias and variance, and analysis of performance

of $\widehat{m}(x)$ can be based on the sizes of these two components. We shall also adopt a commonly-used global measure of closeness between m and $\widehat{m}$, mean integrated squared error (MISE), especially in numerical studies in latter chapters. This is related to MSE by $\text{MISE}\{\widehat{m}(\cdot)\} = \int \text{MSE}\{\widehat{m}(x)\}\, dx$. Unless otherwise specified, the term *rate of convergence* will mean pointwise optimal convergence rate in MSE sense throughout this chapter.

The first kernel estimator that we shall introduce is the *Nadaraya-Watson* estimator $\widehat{m}_{NW}$ (Nadaraya, 1964; Watson, 1964), which is based on a local constant approximation of m. For each x, $\widehat{m}_{NW}(x)$ is defined as the minimiser of

$$\sum_{i=1}^{n} \{Y_i - \widehat{m}_{NW}(x)\}^2 K_h\Big(\frac{X_i - x}{h}\Big). \tag{1.1}$$

Here, K is called the *kernel* function, $K_h(\cdot) = h^{-1}K(\cdot/h)$, and h is known as the *bandwidth*. The function K is usually bounded, continuous, symmetric about 0 and satisfies $\int K = 1$. On minimising (1.1), the Nadaraya-Watson estimator can be given explicitly as

$$\widehat{m}_{NW}(x) = \sum_{i=1}^{n} \Big\{ K_h(X_i - x) \Big/ \sum_{j=1}^{n} K_h(X_j - x) \Big\} Y_i. \tag{1.2}$$

It is clear from (1.2) that $\widehat{m}_{NW}(x)$ is a local weighted average of the Y_i's whose weights $\{K_h(X_i - x)/\sum_{j=1}^{n} K_h(X_j - x)\}_{i=1,\ldots,n}$ are determined by the kernel function K, and hence the name *kernel estimator*. The bandwidth, h, also known as the smoothing parameter, controls the amount of smoothing of the estimator. Loosely speaking, choosing the bandwidth too large results in an over-smoothed estimate with large bias; on the other hand, choosing the bandwidth too small results in an under-smoothed estimate and large variance. For practical applications, the choice of smoothing parameter is a very important issue since it can crucially affect the quality of the estimator. The selection of a suitable bandwidth for kernel estimators (by data-driven means) has been the subject of a number of papers, see for example, Härdle, Hall and Marron (1988, 1992) and Härdle and Marron (1995). By comparison, the selection of a kernel is less influential, and the decision is mostly made on grounds such as computational efficiency. The monograph by Wand and Jones (1995) gives a detailed discussion of kernels.

Bias and variance of $\widehat{m}_{NW}$ admit the following asymptotic approximations:

$$E\{\widehat{m}_{NW}(x) \,|\, X_1, \ldots, X_n\} - m(x) = \frac{1}{2}\kappa_1 h^2 \left\{ m''(x) + \frac{2m'(x)\, f'_X(x)}{f_X(x)} \right\} \{1 + o_p(1)\}, \quad (1.3)$$

$$\operatorname{var}\{\widehat{m}_{NW}(x) \,|\, X_1, \ldots, X_n\} = \frac{\kappa_2 v(x)}{nh\, f_X(x)} \{1 + o_p(1)\}, \quad (1.4)$$

where $\kappa_1 = \int t^2 K(t)\, dt$ and $\kappa_2 = \int K^2$. The deficiencies of $\widehat{m}_{NW}$ are clear from the bias expansion (1.3) (Chu and Marron, 1991; Fan, 1992). First, the estimator is biased even when estimating linear functions $m(x) = \alpha + \beta x$, due to the presence of the term $m'(x) f'_X(x)/f_X(x)$. Large $|\beta|$, or equivalently, large $|m'(x)|$ will typically inflate the bias. Secondly, for non-uniform design where $|f'_X(x)/f_X(x)|$ is large, the bias of $\widehat{m}_{NW}$ is also large. Thus, the Nadaraya-Watson estimator is not adequately design-adaptive.

A better estimator which improves on the bias deficiencies of $\widehat{m}_{NW}$ is the *Gasser-Müller* estimator (Gasser and Müller, 1979), which is given by

$$\widehat{m}_{GM}(x) = \sum_{i=1}^{n} \left\{ \int_{r_i}^{r_{i+1}} K_h(t - x)\, dt \right\} Y_{[i]}, \quad (1.5)$$

where $\{(X_{(i)}, Y_{[i]})\}_{i=1,\ldots,n}$ is an ordered sample with ascending X_i's, $r_0 = -\infty$, $r_{n+1} = +\infty$ and $r_i = (X_{(i)} + X_{(i+1)})/2$. This approach is based on the approximation that $\int m(t)\, K_h(t - x)\, dt$ should be close to $m(x)$ as $h \to 0$, and is related to the convolution smoothing introduced by Clark (1977). Clark suggested convolving a piecewise-linear estimator g with a kernel function K, and proposed the estimator

$$\widehat{m}_{CL}(x) = \int g(t)\, K_h(t - x)\, dt, \quad (1.6)$$

where g is simply a first-order interpolating spline defined by

$$g(t) = \begin{cases} Y_{[1]} & \text{for } t \le X_{(1)}, \\ Y_{[i]} + \frac{Y_{[i+1]} - Y_{[i]}}{X_{(i+1)} - X_{(i)}} (t - X_{(i)}) & \text{for } X_{(i)} \le t \le X_{(i+1)} \ (i = 1, \ldots, n-1), \\ Y_{[n]} & \text{for } t \ge X_{(n)}. \end{cases}$$

If the explanatory variables are equally-spaced on $[0, 1]$, it may be shown that the estimators $\widehat{m}_{GM}$ and $\widehat{m}_{CL}$ are asymptotically equivalent (see for example, Härdle,

1990). The conditional bias and variance of the Gasser-Müller estimator are obtained as

$$E\{\widehat{m}_{GM}(x) \mid X_1, \ldots, X_n\} - m(x) = \tfrac{1}{2}\kappa_1 m''(x)\, h^2 \{1 + o_p(1)\}, \quad (1.7)$$

$$\text{var}\{\widehat{m}_{GM}(x) \mid X_1, \ldots, X_n\} = \tfrac{3}{2}\{nh\, f_X(x)\}^{-1} \kappa_2 v(x) \{1 + o_p(1)\}. \quad (1.8)$$

The bias of the Gasser-Müller estimator, which is independent of the design density f_X and depends only on the curvature of m, has a simpler representation compared to that of the Nadaraya-Watson estimator. It is also easier to interpret, since for large local curvature, $|\, m''(x)\,|$ tends to be large and we should expect more bias to be introduced. The asymptotic variance of $\widehat{m}_{GM}$, however, is 1.5 times that of $\widehat{m}_{NW}$ in the random design model (compare (1.4) and (1.8)). Seifert and Gasser (1996b) used a pictorial illustration to demonstrate why the variance is inflated: if three design points are close together, the middle point receives much less weight compared with the other two points since the weights of the response variables are proportional to the areas under the kernel function between averages of subsequent design points. The Gasser-Müller estimator assigns fluctuating weights to the response variables and increases variability. Several methods have been proposed to alleviate this problem; they include works by Herrmann (1996) and Hall and Turlach (1997a). These methods focus on choices of r_i's (at (1.5)) that reduce the variability of the weights.

1.3 Local Polynomial Fitting

The local polynomial regression technique has been in use for some time in smoothing time series data (Macauley, 1931), and was reviewed systematically by Stone (1977), Cleveland (1979) and Tsybakov (1986). More recent work includes Fan (1992, 1993), Hastie and Loader (1993) and Ruppert and Wand (1994). The revival of interests in local polynomial method is mainly attributed to its favourable sampling properties, which we shall detail in this section. The idea behind the local polynomial kernel estimator, $\widehat{m}_{LP}(x)$, is to approximate $m(x)$ by fitting a p-th degree polynomial locally to the data using weighted least squares around the point x. The weights are, again, determined via a kernel function. Local smoothness of m implies that it can be expanded in a Taylor series and approximated locally by a polynomial.

Specifically, $\widehat{m}_{LP}(x)$ is given by $\hat{\beta}_0$, where $\hat{\beta} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)^T$ is chosen to minimise

$$\sum_{i=1}^{n} \left\{ Y_i - \sum_{j=0}^{p} \beta_j \, (X_i - x)^j \right\}^2 K_h(X_i - x) \, . \tag{1.9}$$

This minimisation problem can also be put in matrix form as follows (Ruppert and Wand, 1994). Denote

$$\mathbf{X} = \begin{pmatrix} 1 & X_1 - x & \ldots & (X_1 - x)^p \\ \vdots & \vdots & & \vdots \\ 1 & X_n - x & \ldots & (X_n - x)^p \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix},$$

and $\mathbf{W} = \text{diag}\,\{K_h(X_1 - x), \ldots, K_h(X_n - x)\}$, the $n \times n$ diagonal matrix of weights. Assuming the invertibility of $\mathbf{X}^T\mathbf{W}\mathbf{X}$, the solution of the least-squares problem (1.9) can be rewritten as

$$\hat{\beta} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y} \, , \tag{1.10}$$

and $\widehat{m}_{LP}(x) = \mathbf{e}^T\hat{\beta}$ where $\mathbf{e}^T = (1, 0, \ldots, 0)$ is a $(n+1) \times 1$ vector. For $p = 0$, the local constant estimator obtained from minimising (1.9) is equivalent to the Nadaraya-Watson estimator (1.2). For $p = 1$, we obtain the local linear kernel estimator

$$\widehat{m}_{LL}(x) = \left(s_0 s_2 - s_1^2\right)^{-1} \sum_{i=1}^{n} \{s_2 - (X_i - x)s_1\} \, K\{(X_i - x)/h\} \, Y_i \, , \tag{1.11}$$

where $s_r = \sum_{i=1}^{n} (X_i - x)^r \, K\{(X_i - x)/h\}$, $r = 0, 1, 2$. An equivalent expression for $\widehat{m}_{LL}(x)$ is

$$\widehat{m}_{LL} = \left(\sum_{i=1}^{n} w_i \, Y_i \right) \Big/ \left(\sum_{i=1}^{n} w_i \right) , \tag{1.12}$$

where $w_i = \{s_2 - (X_i - x)\, s_1\} \, K\{(X_i - x)/h\}$. The theoretical properties of local linear fitting and polynomials of other orders have been well-studied (Fan, 1992, 1993; Ruppert and Wand, 1994; Fan *et al.*, 1997). We shall only detail properties in the local linear case. The conditional bias and variance of $\widehat{m}_{LL}$ are

$$E\left\{\widehat{m}_{LL}(x) \,|\, X_1, \ldots, X_n\right\} = \tfrac{1}{2} \kappa_1 m''(x) \, h^2 \left\{1 + o_p(1)\right\} , \tag{1.13}$$

$$\text{var}\left\{\widehat{m}_{LL}(x) \,|\, X_1, \ldots, X_n\right\} = \frac{\kappa_2 v(x)}{n h \, f_X(x)} \left\{1 + o_p(1)\right\} . \tag{1.14}$$

Fan (1993) showed that the globally optimal bandwidth with respect to integrated conditional mean squared error is

$$h_{opt} = \left\{ \frac{\kappa_2 \int f_X(x)^{-1} v(x)\, dx}{\kappa_1^2 \int m''(x)^2\, dx} \right\}^{1/5} n^{-1/5}, \tag{1.15}$$

in the sense of asymptotically minimising MISE. Both the local linear estimator and the Gasser-Müller estimator have the same asymptotic bias, but the asymptotic variance is smaller for the local linear smoother and is the same as the Nadaraya-Watson estimator in the random design setting. Hence, $\widehat{m}_{LL}(x)$ is superior to the two kernel estimators introduced in the last section, in terms of bias and variance. It is worth mentioning that the degree of the local polynomial, p, determines the size of bias of $\widehat{m}_{LP}(x)$ which decreases as p increases (Ruppert and Wand, 1994). However, the practical gains of high degree fits are doubtful for three reasons: (i) it is computationally costly to solve the minimisation problem (1.9) for large p, (ii) the inversion of $\mathbf{X}^T\mathbf{W}\mathbf{X}$ in (1.10) may create numerical instability in regions with sparse design, and (iii) the variance of $\widehat{m}_{LP}$ is inflated for higher degree fits and a large sample may be needed for practical improvements. For these reasons, one rarely uses local polynomial fits with $p > 3$. See Section 1.4 for more discussion of higher order polynomial fits.

The advantages that $\widehat{m}_{LL}$ offers are more than merely those mentioned above. When the design density f_X has bounded support, say on the closed interval $[a, b]$, a regression smoother using compactly supported kernel normally behaves differently when it reaches a boundary and has slower rate of convergence. We call points lying in the interval $[a + h, b - h]$ *interior* points, and those lying outside this interval *boundary* points. For the Nadaraya-Watson and the Gasser-Müller estimators, bias increases by an order of magnitude to $O(h)$ in estimating boundary points, hence optimal MSE inflates from the usual order $n^{-4/5}$ to $n^{-2/3}$ (Rice, 1984; Gasser and Müller, 1989). While this is only a theoretical result, the boundary effect is, in practice, quite noticeable (Hastie and Loader, 1993). To cope with boundary effects, a popular method is to employ special boundary kernels (Müller, 1984, 1991; Jones, 1993) which typically have the form

$$\overline{K}(t - x) = K(t - x)\{\alpha + \beta(t - x)\},$$

where α and β are determined by moment conditions. Although this method solves the problem of boundary effects, it offers no intuitive interpretation and is arguably

too artificial. Other methods for correcting boundary effects include extrapolation methods (Rice, 1984) and reflection methods (Hall and Wehrly, 1991).

Local linear regression, on the other hand, requires no modification when estimating the boundary (Fan and Gijbels, 1992). Bias and variance at the boundary remain automatically the same order as in the interior. Indeed, since local linear approximation is used in a smaller interval, the bias at a boundary point is smaller than that at an interior point. On the other hand, variance increases at a boundary point since fewer data points lie in the interval. Note that this automatic boundary bias correction is only available for local polynomials of *odd* degree (Ruppert and Wand, 1994). Thus, in data analysis, one normally uses local linear or cubic smoothers.

Another attraction of the local linear smoother comes from a more mathematical viewpoint, *minimax risk* analysis. This gives a measure of how well one estimator performs compared with another under specific functional criteria on the class of estimators and the underlying regression function. Define a *linear smoother* as:

$$\widehat{m}_L(x) = \sum_{i=1}^{n} W_i(x, X_1, \ldots, X_n)\, Y_i\,.$$

The Nadaraya-Watson estimator, the Gasser-Müller estimator and the local linear estimator are clearly linear smoothers from this definition. Denote

$$\mathcal{C}_2 = \{m : |\, m(x) - m(x_0) - m'(x_0)(x - x_0)\,| \le C\,(x - x_0)^2/2\,\}\,,$$

where x_0 is an interior point. Assume also the following conditions:

(i) $v(\cdot)$ is continuous at the point x_0,

(ii) $f_X(\cdot)$ is continuous at the point x_0 with $f_X(x_0) > 0$.

The *linear minimax risk* is defined as

$$R_L(n, \mathcal{C}_2) = \inf_{\widehat{m}_L \text{ linear}} \sup_{m \in \mathcal{C}_2} E\Big[\{\widehat{m}_L(x_0) - m(x_0)\}^2 \,\Big|\, X_1, \ldots, X_n\Big]\,,$$

and the best linear smoother is the one which achieves this linear minimax risk. Fan (1992) showed that the local linear estimator $\widehat{m}_{LL}$ with the *Epanechnikov* kernel, K_e, and bandwidth, h_0 given by

$$K_e(t) = \frac{3}{4}(1 - t^2)\, I\{|\, t\,| \le 1\} \quad \text{and} \quad h_0 = \left\{\frac{15\, v(x_0)^2}{f_X(x_0) C^2 n}\right\}^{1/5}$$

achieves the linear minimax risk. In other words, the *linear minimax efficiency* of $\widehat{m}_{LL}$, defined by

$$\left(\frac{R_L(n, \mathcal{C}_2)}{\sup_{m \in \mathcal{C}_2} E\left[\{\widehat{m}_{LL}(x_0) - m(x_0)\}^2 \middle| X_1, \ldots, X_n\right]} \right)^{5/4}, \tag{1.16}$$

is 100% among all linear smoothers, in an asymptotic sense. Note too that the Nadaraya-Watson estimator has asymptotic linear minimax efficiency 0 since its bias (1.3) depends on the derivative $m'(x_0)$ and its maximal risk is infinite over the class $\mathcal{C}_2$. The Gasser-Müller estimator, with a larger variance component compared with $\widehat{m}_{LL}$, is only 66.7% as efficient as $\widehat{m}_{LL}$. Fan (1993) extended this result and proved that on imposing additional restrictions on the joint density f, the marginal density f_X and the conditional variance v, minimax efficiency (defined as in (1.14) but dropping the constraint in $R_L(n, \mathcal{C}_2)$ that $\widehat{m}_L$ is a linear smoother) remains at 89.4% among *all* estimators.

Assume now that the design density has bounded support. Do the appealing minimax rate properties of $\widehat{m}_{LL}$ extend to estimating boundary points? The answer is affirmative. Cheng, Fan and Marron (1993) showed that the local linear regression estimator achieves 94.4% linear minimax efficiency in estimating the left or right boundary point. Thus, $\widehat{m}_{LL}$ is nearly optimal in estimating the boundary among all linear smoothers. Results of minimax efficiency on local polynomials of other degrees, and on estimating derivatives, are discussed in detail by Fan *et al.* (1997).

1.4 Bias-Reduction Methods

In this section we shall review bias-reduction techniques applicable to general kernel regression methods. By reducing the order of magnitude of bias, one obtains rates of convergence better than the usual $n^{-4/5}$. In the next chapter, we shall introduce a new bias-reduction method in local linear smoothing.

Higher-order kernels were noted by Bartlett (1963) in the context of probability density estimation. We call K a *j-th order kernel* if it satisfies

$$\int K = 1\,, \quad \int t^i K(t)\,dt = 0 \text{ for } i = 1, \ldots, j-1, \text{ and } \int t^j K(t)\,dt \neq 0\,.$$

In other words, a j-th order kernel integrates to 1 and has vanishing first $(j-1)$-th

moments. A general j-th order kernel smoother has the form

$$\widehat{m}(x) = \frac{1}{nh} \sum_{i=1}^{n} K_{(j)}\Big(\frac{X_i - x}{h}\Big) Y_i \,,$$

where $K_{(j)}$ denotes a kernel of order j. To see why higher-order kernel techniques can reduce bias, we note that the conditional expectation of $\widehat{m}(x)$ is

$$E\big\{\widehat{m}(x) \big| X_1, \ldots, X_n\big\} = \int K_{(j)}(u)\, m(x + uh)\, du \,.$$

Assuming sufficient regularity conditions, $m(x + uh)$ may be expanded as a Taylor series about x and the moment conditions on $K_{(j)}$ ensure that the bias of $\widehat{m}(x)$ is of order h^j. Notice that higher-order kernels take negative values, and that the overall performance of $\widehat{m}(x)$ may be undesirably affected. It is more difficult to interpret the resulting estimator when negative weights are assigned to some of the Y_i's. In the related context of kernel density estimation, use of higher-order kernels may even result in a negative density estimate.

As mentioned in the previous section, the degree of the local polynomial determines the order of bias of the estimator $\widehat{m}_{LP}$. Ruppert and Wand (1994) showed that for local pth-degree polynomial fits, conditional bias admits the following formulae:

$$\begin{aligned} &E\big\{\widehat{m}_{LP}(x) \,\big|\, X_1, \ldots, X_n\big\} \\ &\quad = h^{p+1} \left\{\frac{m^{(p+1)}}{(p+1)!}\right\} \left\{\int u^{p+1} K_{[p]}(u)\, du\right\} \big\{1 + o_p(1)\big\} \qquad \text{if } p \text{ is odd}\,, \\ &\quad = h^{p+2} \left\{\frac{m^{(p+1)}(x)\, f_X'(x)}{(p+1)!\, f(x)} + \frac{m^{(p+2)}(x)}{(p+2)!}\right\} \left\{\int u^{p+2} K_{[p]}(u)\, du\right\} \big\{1 + o_p(1)\big\} \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } p \text{ is even}\,, \quad (1.17) \end{aligned}$$

where $K_{[p]}(u) = \{\, |\mathbf{M}_p(u)| \,/\, |\mathbf{N}_p| \,\}\, K(u)$, $\mathbf{N}_p$ is a $(p+1) \times (p+1)$ matrix having (i,j)-th entry equal to $\int u^{i+j-2}\, K(u)\, du$, and $\mathbf{M}_p(u)$ is the same as $\mathbf{N}_p$ but with the first column replaced by $(1, u, \ldots, u^p)^T$. Lejeune and Sarda (1992) showed that $K_{[p]}$ is a kernel of order $p+1$ for p odd, and is of order $p+2$ for p even. Indeed, local constant and linear fits resemble second-order kernel estimation, and local quadratic and cubic fits resemble fourth-order kernel estimation. Fan and Gijbels (1995) proposed a data-driven variable-order selection procedure in which the order of fit is chosen adaptively.

Another bias-reduction method was given by Härdle (1986) using a *jackknife* technique. This approach is very similar to that proposed by Rice (1984), who combined two kernel estimators to reduce boundary bias, which was motivated by Richardson extrapolation. Let $\widehat{m}_{h_j}(x)$ be a kernel smoother with bandwidth h_j which admits asymptotic bias $C(K)\,m''(x)\,h_j^2$ (see (1.7) and (1.13)) for $j = 1, 2$, and $C(K)$ is a constant which depends only on K. The jackknife estimator is given by

$$\widehat{m}_J(x) = (1-\omega)^{-1}\{\widehat{m}_{h_1}(x) - \omega\,\widehat{m}_{h_2}(x)\},$$

and has asymptotic bias given by $b(x) = (1-\omega)^{-1}\,(h_1^2 - \omega\,h_2^2)\,C(K)\,m''(x)$, provided $\omega \neq 1$. The choice of ω is crucial here, since $b(x) = 0$ if ω is taken to be h_1^2/h_2^2. Hence, the bias of $\widehat{m}_J(x)$ is reduced compared with $\widehat{m}_{h_1}(x)$ or $\widehat{m}_{h_2}(x)$. Note that $\widehat{m}_J(x)$ is equivalent to the kernel smoother based on the kernel $L(u,\omega) = (1 - \omega)\{K(u) - \omega^{3/2}\,K(\omega^{1/2}u)\}$, with $\omega = h_1^2/h_2^2$, where it may be easily shown that $L(u,\omega)$ is a fourth-order kernel. Thus, jackknifing is essentially a high-order kernel method. A comparison of efficiency of the jackknifed kernel smoother with respect to ordinary kernel estimators can be found in Härdle (1986). In practice, one has to jointly select h_1 and ω (or equivalently, h_1 and h_2), and the performance of $\widehat{m}_J$ seems to be fairly sensitive to the choice of ω.

Variable bandwidth bias-reduction methods were introduced by Breiman, Meisel and Purcell (1977) and Abramson (1982) in the context of density estimation. Instead of choosing a constant bandwidth over the entire range of inference, the bandwidth is allowed to vary and depends on the data. In nonparametric regression, a kernel estimator of $m(x)$, with variable bandwidths $h_1/\alpha_1(X_i)$ in the numerator and $h_2/\alpha_2(X_i)$ in the denominator, is given by

$$\widehat{m}_{VB}(x) = \frac{(nh_1)^{-1}\sum_{i=1}^n \alpha_1(X_i)\,K\{(x - X_i)\,\alpha_1(X_i)/h_1\}\,Y_i}{(nh_2)^{-1}\sum_{i=1}^n \alpha_2(X_i)\,K\{(x - X_i)\,\alpha_2(X_i)/h_2\}},$$

where K is assumed to be a second-order kernel. Hall (1990) studied the bias of a variable bandwidth estimator in very general settings and showed explicitly how to determine appropriate α_1 and α_2 for minimising bias. He recommended, theoretically, taking $\alpha_1 = |\,mf_X\,|^{1/2}$ and $\alpha_2 = |\,f_X\,|^{1/2}$. If one chooses $h_1 = h_2$, then bias of $\widehat{m}_{VB}(x)$ has size h_1^4. In practice, $\alpha_1(x)$ and $\alpha_2(x)$ are estimated by

$$\widehat{\alpha}_1(x) = \left|\frac{1}{nh_3}\sum_{i=1}^n K\Big(\frac{X_i - x}{h_3}\Big)\,Y_i\right|^{1/2} \quad \text{and} \quad \widehat{\alpha}_2(x) = \left|\frac{1}{nh_4}\sum_{i=1}^n K\Big(\frac{X_i - x}{h_4}\Big)\right|^{1/2}$$

respectively, where h_3, h_4 are bandwidths of size $n^{-1/5}$. The "adaptive form" estimator, $\widetilde{m}_{VB}$, defined by substituting $\widehat{\alpha}_j$ for α_j in $\widehat{m}_{VB}$, preserves the bias-reduction qualities of $\widehat{m}_{VB}$ without appreciably affecting variance. Though this method is capable of reducing bias, its complicated nature makes it less appealing in practice.

The multiplicative bias-reduced estimator (MBRE) for nonparametric regression was proposed by Linton and Nielsen (1994). It has mean squared error of order $n^{-8/9}$. The idea is relatively straightforward. An initial smooth, say $\widetilde{m}(x)$, is obtained. Write $m(x) = \alpha(x)\widetilde{m}(x)$ where $\alpha(x) = m(x)/\widetilde{m}(x)$. The MBRE is defined as $\widehat{m}_{MB}(x) = \widetilde{\alpha}(x)\widetilde{m}(x)$, and $\widetilde{\alpha}(x)$ is an estimate of $\alpha(x)$. By suitably choosing $\widetilde{\alpha}(x)$ it is possible to reduce bias from size h^2 to h^4. Linton and Nielsen (1994) treated only the case for equispaced design. Jones, Linton and Nielsen (1995) proposed a more generally-applicable version by employing the local linear estimator $\widehat{m}_{LL}(x)$ as the initial smooth, and multiplying this by a local linear regression of $Y_i/\widehat{m}_{LL}(X_i)$ on X_i. To be explicit, they defined their estimator as

$$\widehat{m}_{MB}(x) = \widehat{m}_{LL}(x) \sum_{i=1}^{n} \frac{\{s_2 - (X_i - x)s_1\} K\{(X_i - x)/h\}}{(s_0 s_2 - s_1^2)} \left\{ \frac{Y_i}{\widehat{m}_{LL}(X_i)} \right\},$$

where s_i's are defined as in (1.11). Jones, Linton and Nielsen (1995) showed that the bias of this estimator is of order h^4, while the variance remains the same order $(nh)^{-1}$.

1.5 Overcoming Sparse Design

The merits of local polynomial methods have already been discussed in Section 1.3. Nevertheless, their theoretical attractions are, on occasions, undermined by their practical performance. To give an explicit example, the curve estimate in Figure 1.1 was constructed using local linear fitting with bandwidth chosen to minimise the asymptotic MISE (see (1.15)). The target was $m(x) = 2\sin(4\pi x)$, and the sample size was $n = 50$ with uniform design density. The estimate is relatively erratic in places where the design is sparse. This may be explained by the fact that the conditional variance of local polynomial methods has no upper bound, and the unconditional variance when using a compact kernel is infinite. In this section we shall look at a few remedies for the sparse design problem, and concentrate mainly on cures for the local linear estimator. Some of the methods are applicable to local polynomials of higher degree, which we shall briefly indicate. Throughout this

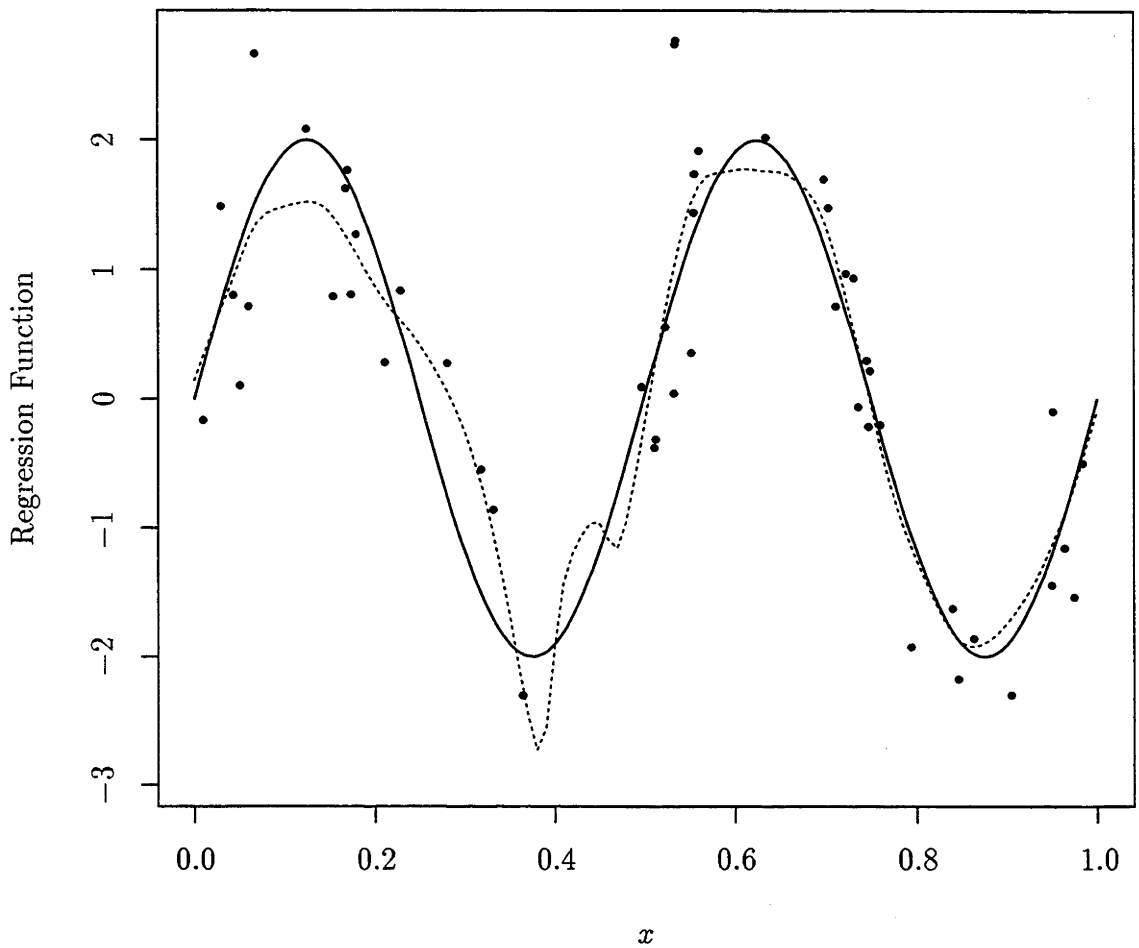


Figure 1.1: Local linear kernel estimate for the regression function $m(x) = 2\sin 4\pi x$ *with sample size* $n = 50$. The solid curve is the true regression function, and the dotted line is the local linear kernel estimate. The Standard Normal kernel is employed, and the bandwidth $h = 0.0335$ is chosen to minimise the asymptotic MISE.

section, the local linear smoother will be denoted by $\widehat{m}$, and we assume the kernel used to construct $\widehat{m}$ is compactly supported on $[-1, 1]$.

Seifert and Gasser (1996a, 1996b) proposed two modifications: (i) by locally increasing the bandwidth in regions of sparse design to allow sufficient data points to be included, and (ii) by incorporating a ridge parameter to stabilise the variance. In the first strategy, let h_0 be a pre-assigned bandwidth for estimating $m(x_0)$. We wish to choose h_1 close to h_0 such that $\widehat{m}(x_0)$ has an acceptable level of variance. Seifert

and Gasser suggested that this can be done in two steps. First, compute the finite-sample variance $\widehat{V}_{h_0}\{\widehat{m}(x_0)|\, X_1,\ldots,X_n\}$ and compare it with some reference level, say $\delta\,\widetilde{V}_{h_0}$, where δ is a pre-specified constant, and $\widetilde{V}_{h_0}$ is the asymptotic variance (1.14) calculated for uniform design. When $\widehat{V}_{h_0}\{\widehat{m}(x_0)|\, X_1,\ldots,X_n\}$ exceeds the reference level $\delta\,\widetilde{V}_{h_0}$, the bandwidth is locally increased. It may be that, however, an increase in bandwidth does not result in a sufficient decrease in variance. In this case, variance-bias consideration comes into play. As the bandwidth increases from h_0 to h_1, the asymptotic squared bias increases in proportion to $(h_1/h_0)^4$. Thus, as a second step, one finds h to minimise

$$\widehat{V}_{h_0}\{\widehat{m}(x_0)|\, X_1,\ldots,X_n\} + \Big(\frac{h}{\sqrt{2}\,h_0}\Big)^4 \widetilde{V}_{h_0}\,. \tag{1.18}$$

Extension to local polynomials of other degrees, as well as further arguments that lead to minimising (1.18), can be found in Seifert and Gasser (1996a). They also showed, through simulation studies, that the choice of $\delta = 1.0$ behaves well.

Another method introduced by Seifert and Gasser (1996a, 1996b) is to incorporate a parameter, known as a "ridge", in the estimator. Recall that in (1.10), calculation of $\hat{\beta}$ involves the inversion of $\mathbf{X}^T\mathbf{W}\mathbf{X}$. In regions of sparse design, this matrix is close to singular or even non-invertible. Ridging guarantees that this matrix is non-singular, through the addition of a positive semidefinite matrix $\mathbf{H}$ such that $\mathbf{H} + \mathbf{X}^T\mathbf{W}\mathbf{X}$ is non-singular; and the ridged estimator is given by

$$\widetilde{\beta} = (\mathbf{H} + \mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}\,.$$

The principle involved in ridging has in fact been adopted by Fan (1993) in proving the optimal performance of the local linear smoother. He defined the local linear smoother (1.12) slightly different by adding a factor of n^{-2} to the denominator :

$$\widetilde{m}(x) = \Big(\sum_{i=1}^{n} w_i\, Y_i\Big) \Big/ \Big(\sum_{i=1}^{n} w_i + n^{-2}\Big)\,. \tag{1.19}$$

This has the effect of avoiding zero in the denominator when there are no design points in the interval $\mathcal{I} = [x-h, x+h]$. Seifert and Gasser proposed several choices of $\mathbf{H}$ under certain smoothness restrictions on the regression function. However, their choice does not seem to remedy the sparse design problem, if there are insufficient design points in $\mathcal{I}$, unless h is increased locally. Moreover, the ridge parameter, which can be regarded as another smoothing parameter, has to be chosen empirically

together with the bandwidth, and the performance of the estimator can be seriously impaired by a poor choice of the ridge parameter. We shall demonstrate this in the numerical section in the next chapter.

Another technique, formulated by Cheng, Hall and Titterington (1997), involves shrinking a local linear smoother towards a general curve estimator, $\bar{m}$. Indeed, ridging described in the above paragraph can be viewed as shrinking $\widehat{m}$ towards 0. Recall that in local linear smoothing, one chooses a and b to minimise

$$\sum_{i=1}^{n} \{Y_i - a - b(X_i - x)\}^2 K\Big(\frac{X_i - x}{h}\Big), \tag{1.20}$$

and $\widehat{m}$ is defined by a. In shrinkage, the expression at (1.20) is generalised, and one looks for a and b to minimise

$$\sum_{i=1}^{n} \{Y_i - a - b(X_i - x)\}^2 K\Big(\frac{X_i - x}{h}\Big) + \epsilon\{\bar{m}(x) - a\}^2, \tag{1.21}$$

where $\epsilon = \epsilon(x) > 0$. The new estimator, $\widehat{m}_S$, is taken as a in the minimisation of (1.21) and can be given explicitly as

$$\widehat{m}_S = \Big(\sum_{i=1}^{n} w_i Y_i + \epsilon s_2 \bar{m}\Big) \Big/ \Big(\sum_{i=1}^{n} w_i + \epsilon s_2\Big), \tag{1.22}$$

where w_i and s_2 are defined as in (1.12). Taking $\bar{m} = 0$ and $\epsilon = (n^2 s_2)^{-1}$ in (1.22) gives the ridged version of the local linear smoother (1.19). Note that $\widehat{m}_S = \widehat{m}$ when $\epsilon = 0$, and $\widehat{m}_S = \bar{m}$ if $\epsilon = \infty$. Cheng, Hall and Titterington (1997) suggested taking $\bar{m}$ to be another local linear smoother, constructed using the same bandwidth but with an infinitely supported kernel to guard against the sparse-design problem. The advantage of this approach is its ability to produce a proper curve estimator even when an excessively large shrinkage parameter is chosen. This was supported both in the theoretical and numerical studies by Cheng, Hall and Titterington. Moreover, $\widehat{m}_S$ can be regarded as a "mixture" of two local linear estimators using compactly and infinitely supported kernels respectively, and enjoys the merits of reduced edge effect and lower mean squared error (if the Epanechnikov kernel is used) from the former, and increased numerical stability from the latter.

Interpolation methods (Hall and Turlach, 1997b) involve imputing pseudo design points to overcome data sparseness problems in local linear smoothing. The rules for determining where to add the pseudo points are simple, and depend only on

the kernel and the bandwidth used to construct the local linear estimator. The methods are applicable to all bandwidths, and are easy to implement. Indeed, using a suitable interpolation rule allows the conditional mean squared error of a local linear smoother to be stated in an unconditional sense.

Assume that the random sample $\{(X_i, Y_i)\}_{i=1,\ldots,n}$ has its predictor variables X_i sorted in ascending order, i.e. $X_1 \leq \ldots \leq X_n$; that the kernel K has support on $[-1, 1]$; and that the X_i's lie in the compact interval $[a, b]$. Put $(X_0, Y_0) = (a, Y_1)$ and $(X_{n+1}, Y_{n+1}) = (b, Y_n)$, and let $S_i = X_{i+1} - X_i$ for $0 \leq i \leq n$. Let $\mathcal{J}_h$ denote the set of indices i such that $a + h \leq X_i \leq X_{i+1} \leq b - h$, where h is the bandwidth used in the local linear estimator. For a given real number r, write m_i for the integer part of $rS_i/(2h)$ if $i \in \mathcal{J}_h$, and for the integer part of rS_i/h otherwise. If $m_i \geq 1$, add m_i equally-spaced pseudo design points to the interval $[X_i, X_{i+1}]$. For each pseudo point, the corresponding Y-value is generated by linear interpolation between the points (X_i, Y_i) and (X_{i+1}, Y_{i+1}). Essentially, the interval $[X_i, X_{i+1}]$ is divided into $m_i + 1$ equal portions of length not exceeding $2h/r$ if $i \in \mathcal{J}_h$, and h/r otherwise. If $m_i = 0$, no pseudo design point is added. This rule ensures that none of the distances between two adjacent design or pseudo points is more that $2h/r$. Equivalently, for each $x \in [a, b]$, the number of those points in the interval $(x - h, x + h)$ is at least equal to the integer part of r.

Figure 1.2 demonstrates the interpolation rule with $r = 3$ using the same example as in Figure 1.1. The Epanechnikov kernel was employed, and the bandwidth was chosen to minimise the asymptotic MISE. The pseudo points are indicated by crosses in the diagram and can be found in those places where the "real" design was sparse. The numerical studies by Hall and Turlach (1997b) showed that the performance of the approach is fairly insensitive to choice of r, and in many cases, has better mean-square performance than the local ridge regression approach proposed by Seifert and Gasser (1996a, 1996b).

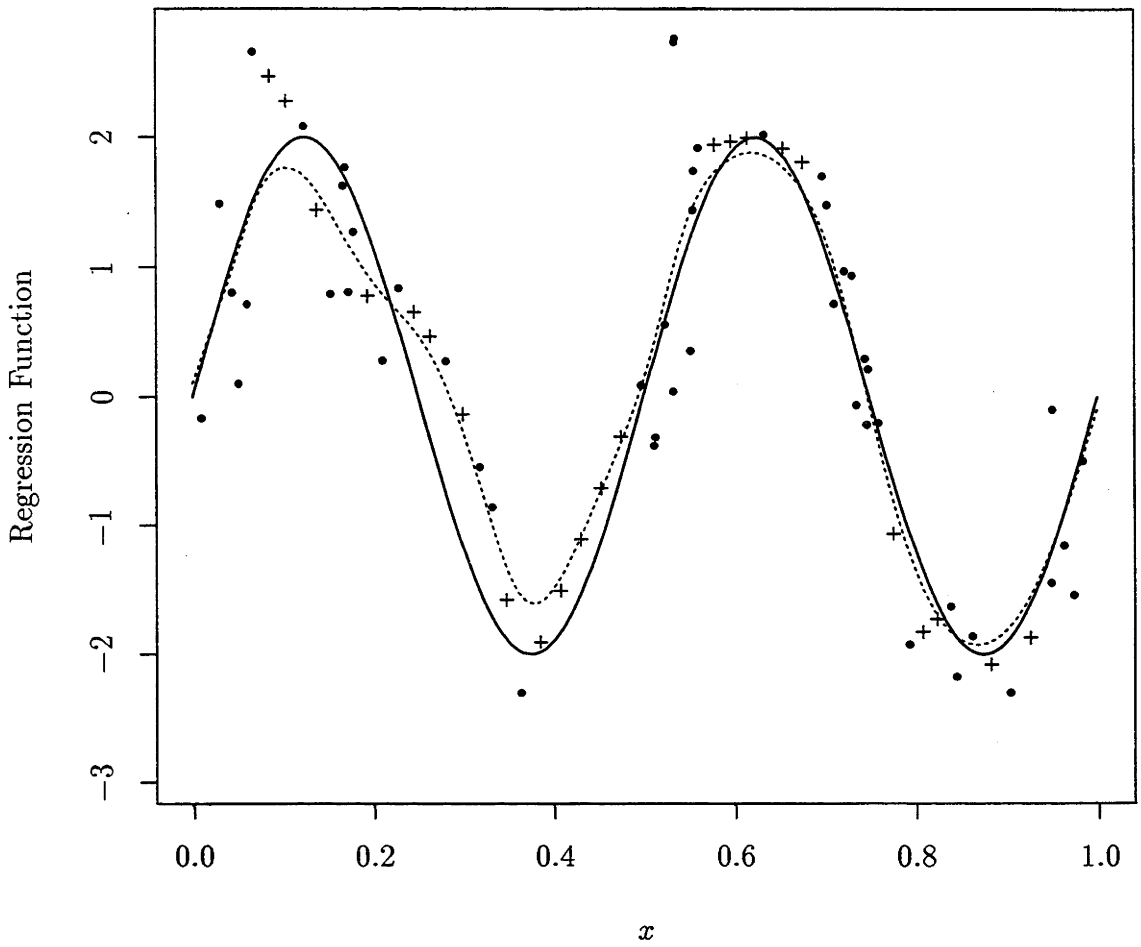


Figure 1.2: Local linear kernel estimate for the regression function $m(x) = 2\sin 4\pi x$ *with sample size* $n = 50$*, using the interpolation rule of Hall and Turlach (1997b).* The solid curve is the true regression function, and the dotted line is the local linear kernel estimate. The Standard Normal kernel is employed, and the bandwidth $h = 0.0335$ is chosen to minimise the asymptotic MISE. We used the same sample as in Figure 1.1, except that pseudo-data points, represented by crosses in the figure, were added to ensure there were at least three points in each interval $(x-h, x+h)$, $x \in [0,1]$.

1.6 Summary

This chapter highlights some favourable properties of local polynomial estimators. The local linear estimator, in particular, enjoys excellent numerical as well as theoretical properties (e.g. Fan, 1993; Hastie and Loader, 1993; Cleveland and Loader, 1996). Among all linear estimators, it is 100% efficient in estimating regression means with two bounded derivatives, and nearly 90% efficient among all estimators in an asymptotic minimax sense (Fan 1993). Widely-used smoothing software is based on local linear methods; see for example Cleveland (1979, 1993), Cleveland and Devlin (1988) and Cleveland and Grosse (1991). We also review bias-reduction methods, all of which reduce bias from the usual order h^2 to h^4. We further look at techniques for overcoming the problem of sparse design in local linear smoothing. In the next chapter we shall see how one can achieve bias reduction by modifying the usual local linear estimator, the mechanism of which may be explained geometrically. We shall demonstrate how the interpolation device developed by Hall and Turlach (1997b) may be easily adapted to our new estimators.

Chapter 2

Bias Reduction

2.1 Introduction

The favourable properties of local linear smoothing were discussed in the last chapter. For regression functions that exhibit a high degree of smoothness, local polynomial methods of higher order are, at least in theory, superior to local linear approaches in reducing bias, as mentioned in Section 1.4. Nevertheless, higher-degree fits require necessarily more elaborate techniques for guarding against data sparseness problems, compared to those for local linear smoothing. For example, to obtain a local cubic estimator, one needs to invert a 4×4 matrix (see (1.10)) and to avoid numerical problems in regions where the design is sparse, one has to ensure that this matrix is not close to being singular.

In this chapter, we shall demonstrate how one may achieve bias reduction by combining two or three linear estimators, obtaining essentially the same optimal performance as the local cubic smoother. The techniques employed by local linear estimators to guard against data sparseness problems may be applied directly to our new estimators. Essentially, our method involves a convex combination of local linear estimators with easily-chosen weights that depend only on the kernel function, and not at all on other unknowns like the regression mean or design density. It has similar spirit to the bias-reduction methods introduced by Schucany and Sommers (1977) and Härdle (1986), who suggested using a linear combination of two kernel estimators of different bandwidths, and has already been discussed in Section 1.4.

Figure 2.1 demonstrates a simple graphical property which motivates our method. The true regression mean is convex, and the standard local linear estimator tends

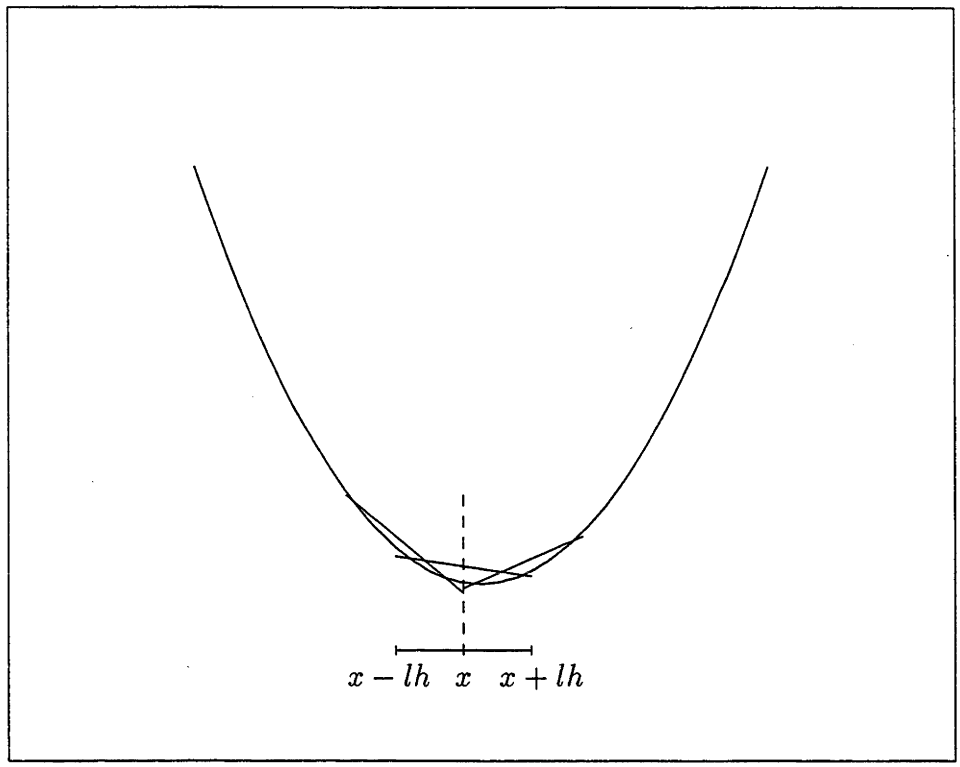


Figure 2.1: Bias reduction via a convex combination of three local linear smoothers. By choosing the weights in an appropriate way, bias contributions from the two asymmetric smooths on either side of the symmetric smooth will cancel those of the latter, resulting in reduction of bias by two orders of magnitude. For a slightly different choice of either of the asymmetric smooths, the line segment will cut the curve at a point whose abscissa is very close to x, and so reduce bias by one order of magnitude.

to be positively biased since it is based on the midpoint of a fitted line segment and that point always lies above the curve. (We assume here that there is no noise, which is appropriate when describing the effect of bias.) The ends of the fitted line segment, however, lie below the curve, and there is potential for employing this asymmetry to cancel out the larger part of bias. By using segments on either side of the original one, as indicated in the figure, and employing an appropriate weighted average of two such estimators and the classical local linear smoother, we may re-

move all the first- and second-order effects of bias. In fact, as we shall show later, this reduces bias by two orders of magnitude compared with standard local linear smoothing. Variance may be reduced, although only by a constant factor, not an order of magnitude. The amount of variance reduction depends on the kernel type.

Using a single local linear estimator calculated in a slightly asymmetric way, bias and variance have the same order as for a local quadratic smoother. This approach may be motivated by considering the construction of the local regression line so as to ensure that the expected value of the place where the line crosses the curve has its abscissa very close to the one at which we wish to estimate the curve. To a significant extent this may be guaranteed without prior knowledge of the curve. The average of two of these smooths, on either side of the point at which we wish to estimate the regression mean, reduces bias by two orders of magnitude. Indeed, this average can be viewed as a limiting form of the estimator described in the paragraph above. We shall term our technique "skewing", which reflects the use of asymmetric methods to improve performance.

This chapter is organised as follows. Section 2.2 introduces a general version of the skewed estimators, and Section 2.3 presents main theoretical properties. Left- and right-skewed estimators are introduced in Section 2.4. Section 2.5 discusses general issues and extensions to our skewing methods. Numerical performance is addressed in Section 2.6. There, we show that a simple interpolation device (Hall and Turlach, 1997b), borrowed from the case of ordinary local linear smoothing, may be used to guard against sparse design.

2.2 General Skewed Estimators

Suppose we observe pairs of random variables $(X_1, Y_1), \ldots, (X_n, Y_n)$ drawn independently from a bivariate distribution. We are interested in estimating the regression mean, $m(x) = E(Y \mid X = x)$, where (X, Y) denotes a generic pair of random variables from the sample. In local linear regression, the line $y(u) = a + b(u - x)$ is fitted by weighted least-squares to the data pairs (X_i, Y_i) for those X_i's which are in the neighbourhood of x. The weights given to individual data points are determined by the kernel function K, which is assumed to be non-negative and symmetric. The

pair (a, b) is obtained by minimising

$$\sum_{i=1}^{n} \{Y_i - a - b(X_i - x)\}^2 K_h(X_i - x),$$

where $K_h(\cdot) = h^{-1} K(\cdot/h)$ and h is a bandwidth. The minimising pair (a, b) depends on x as well as on the data, and is denoted by $\{\hat{a}(x), \hat{b}(x)\}$. Elementary calculus shows that the minimisers are

$$\hat{a}(x) = \frac{r_0(x)\, s_2(x) - r_1(x)\, s_1(x)}{s_0(x)\, s_2(x) - s_1(x)^2}, \quad \hat{b}(x) = \frac{r_1(x)\, s_0(x) - r_0(x)\, s_1(x)}{s_0(x)\, s_2(x) - s_1(x)^2},$$

where $r_l(x) = \sum_{i=1}^{n} (X_i - x)^l K_h(X_i - x)\, Y_i$ and $s_l(x) = \sum_{i=1}^{n} (X_i - x)^l K_h(X_i - x)$, $l = 0, 1, 2, \ldots$.

The estimator of the line $y = y(u)$ is $\widehat{m}(u|x) = \hat{a}(x) + \hat{b}(x)\,(u - x)$, with formula

$$\widehat{m}(u|x) = \frac{r_0(x)\, s_2(x) - r_1(x)\, s_1(x) + \{r_1(x)\, s_0(x) - r_0(x)\, s_1(x)\}\,(u - x)}{s_0(x)\, s_2(x) - s_1(x)^2}. \quad (2.1)$$

The standard approach to local linear regression involves fitting a straight line segment whose midpoint is directly above the point x at which we wish to estimate the curve. Putting $u = x$ in (2.1) gives the usual local linear estimator $\widehat{m}(x) = \widehat{m}(x|x) = \hat{a}(x)$, which has conditional bias of size h^2 and conditional variance of size $(nh)^{-1}$ (Fan, 1993). Skewing involves fitting the straight line segment in an asymmetric manner, with its centre a little to the left or right of x. A general skewed estimator $\widetilde{m}$ is a convex combination of three local linear smoothers

$$\widetilde{m}(x) = \frac{\lambda_1 \widehat{m}(x|x + l_1 h) + \widehat{m}(x|x) + \lambda_2 \widehat{m}(x|x + l_2 h)}{\lambda_1 + 1 + \lambda_2}, \quad (2.2)$$

where $\lambda_1, \lambda_2 > 0$ are weights, $l_1 < 0$ and $l_2 > 0$. Versions of (2.2) will be described in Section 2.4. Intuition suggests that we take $\lambda_1 = \lambda_2 = \lambda$ and $l_1 = -l_2 = l$, say, so as to enhance the symmetrical structure of $\widetilde{m}(x)$ and reduce bias. In fact, this choice is necessary if we want to reduce bias by two orders of magnitude. Theorem 2.1 in the next section shows this explicitly. The theorem also shows that the parameters λ and l are related by a simple relation which depends only on the kernel function.

2.3 Theoretical Properties

We shall derive the asymptotic conditional bias and conditional variance of $\widetilde{m}$ in this section. We suppose that the design variables X_i come from a continuous distribution with density f. The j-th moment of K, $\int u^j K(u)\, du$, is denoted by κ_j.

Theorem 2.1 *Assume that m has four bounded, continuous derivatives in a neighbourhood of x; that f has two bounded, continuous derivatives there and $f(x) > 0$; that the kernel K is non-negative, bounded, symmetric and compactly supported, with $\int K = 1$; and that $h = h(n) \to 0$ and $nh \to \infty$. Take $\lambda_1 = \lambda_2 = \lambda > 0$ and $l_1 = -l_2 = l(\lambda)$, where*

$$l(\lambda) = \{(1 + 2\lambda)\,\kappa_2/(2\lambda)\}^{1/2}. \tag{2.3}$$

Then the bias of $\widetilde{m}$ is given by

$$E\{\widetilde{m}(x) - m(x)|X_1, \ldots, X_n\} = B(x)\,h^4 + o_p\{h^4 + (nh)^{-1/2}\},$$

where

$$B(x) = \{16f(x)\}^{-1}\left[2\{2f''(x)\,m''(x) + 4f'(x)\,m'''(x) + f(x)\,m^{(iv)}(x)\}\right.$$
$$\left.\times(\kappa_2^2 - \kappa_4) - \lambda^{-1}\kappa_2^2 f(x)\,m^{(iv)}(x)\right].$$

Remark 2.1. The theorem actually holds under weaker symmetry conditions than those imposed on K. In particular, we require only $\kappa_1 = \kappa_3 = 0$, not symmetry. Hence we shall retain terms in κ_5 in our proof below.

Remark 2.2. It is, in fact, not necessary to assume that a skewed estimator has the form $\widehat{m}(x|x \pm k)$ where $k = lh$. The fact that k is of size h may be deduced directly from the proof.

Proof of Theorem 2.1. Put $\hat{\mu}(u|x) = E\{\widehat{m}(u|x)|\,X_1, \ldots, X_n\}$, which we may expand as

$$\hat{\mu}(u|x) = m(x) + (u - x)m'(x) + \{s_0(x)\,s_2(x) - s_1(x)^2\}^{-1}\{Q(x) + R(x)\}, \tag{2.4}$$

where $Q(x)$ equals

$$\begin{aligned}
&\tfrac{1}{2}\,m''(x)\left[\{s_2(x)^2 - s_3(x)\,s_1(x)\} + (u - x)\{s_3(x)\,s_0(x) - s_2(x)\,s_1(x)\}\right] \\
&+\tfrac{1}{6}\,m'''(x)\left[\{s_3(x)s_2(x) - s_4(x)\,s_1(x)\} + (u - x)\{s_4(x)s_0(x) - s_3(x)s_1(x)\}\right] \\
&+\tfrac{1}{24}\,m^{(iv)}(x)\left[\{s_4(x)\,s_2(x) - s_5(x)\,s_1(x)\} + (u - x)\{s_5(x)s_0(x) - s_4(x)s_1(x)\}\right],
\end{aligned} \tag{2.5}$$

and $R(x)$ may be expressed concisely using an exact formula for the reminder in Taylor's theorem. It may be easily shown that

$$(nh^{l+1})^{-1}\,s_l(x) = \int u^l\,K(u)\,f(x + uh)\,du + (nh)^{-1/2}Z_l,$$

where Z_l is a random variable which is asymptotically normally distributed, with mean 0 and variance $f(x)\int u^{2l}K(u)^2\,du$. It follows from Taylor expansion that

$$\begin{aligned}(nh^{l+1})^{-1}s_l(x) = \kappa_l\, f(x) + \kappa_{l+1}\, f'(x)\, h \\ +\tfrac{1}{2}\kappa_{l+2}\, f''(x)\, h^2 + o(h^2) + O_p\{(nh)^{-1/2}\}\,. \qquad (2.6)\end{aligned}$$

Let γ_l and δ_l denote generic positive numbers which equal $o(h^l)+O\{(nh)^{-1/2}\}$ and $o(h^l)+O\{h^2\,(nh)^{-1/2}\}$ respectively, and define $r_{kl}(x) = s_k(x)s_l(x) - s_{k+l-1}(x)s_1(x)$. Then, using (2.6), we may obtain:

$$\begin{aligned}
n^2h^4\, r_{02}(x)^{-1} &= \frac{1}{f(x)^2\kappa_2}\left[1 - \frac{1}{2}\left\{\frac{f''(x)}{f(x)}\left(\frac{\kappa_2^2+\kappa_4}{\kappa_2}\right)\right.\right. \\
&\qquad \left.\left. -2\left(\frac{f'(x)}{f(x)}\right)^2\kappa_2\right\}h^2\right] + o_p(\gamma_2)\,, \\
n^{-2}h^{-6}\, r_{22}(x) &= \{f(x)\,\kappa_2\}^2 + \left\{f''(x)\,f(x) - f'(x)^2\right\}\kappa_2\kappa_4\, h^2 + o_p(\gamma_2)\,, \\
n^{-2}h^{-5}\, r_{30}(x) &= f(x)\,f'(x)\,(\kappa_4 - \kappa_2^2)\,h^2 + \frac{1}{2}f(x)\,f''(x)\,\kappa_5\,h^2 + o_p(\gamma_2)\,, \\
n^{-2}h^{-7}\, r_{32}(x) &= \frac{1}{2}\left\{f(x)\,f''(x) - 2\,f'(x)^2\right\}\kappa_2\kappa_5\,h^2 + o_p(\gamma_2)\,, \\
n^{-2}h^{-6}\, r_{40}(x) &= f(x)^2\,\kappa_4 + f(x)\,f'(x)\,\kappa_5\,h + o_p(\gamma_1)\,, \\
n^{-2}h^{-8}\, r_{42}(x) &= f(x)^2\,\kappa_2\kappa_4 + o_p(\gamma_0)\,, \\
n^{-2}h^{-7}\, r_{50}(x) &= f(x)^2\,\kappa_5 + o_p(\gamma_0)\,.
\end{aligned}$$

Defining $t_{kl}(x) = r_{02}(x)^{-1}r_{kl}(x)$, we may deduce the following formulae:

$$\begin{aligned}
t_{22} &= \kappa_2 h^2 + \left\{\frac{f''(x)}{2f(x)} - \left(\frac{f'(x)}{f(x)}\right)^2\right\}(\kappa_4 - \kappa_2^2)\,h^4 + O_p(\delta_4)\,, \\
t_{30} &= \frac{f'(x)\,(\kappa_4 - \kappa_2^2)}{f(x)\,\kappa_2}\,h^2 + \frac{f''(x)\,\kappa_5}{2f(x)\,\kappa_2}\,h^3 + O_p(\delta_3)\,, \\
t_{32} &= O_p(\delta_4)\,, \qquad t_{40} = \frac{\kappa_4}{\kappa_2}\,h^2 + \frac{f'(x)\,\kappa_5}{f(x)\kappa_2}\,h^3 + O_p(\delta_3)\,, \\
t_{42} &= \kappa_4 h^4 + O_p(\delta_4)\,, \quad t_{50} = \frac{\kappa_5}{\kappa_2}\,h^3 + O_p(\delta_3)\,.
\end{aligned}$$

Substituting these results into (2.5), substituting the expansion of $Q(x)$ into (2.4), and developing a similar but more crude approximation to the remainder $R(x)$, we

may prove that for any fixed l,

$$\begin{aligned}\hat{\mu}(x|x+lh) &= m(x) + \tfrac{1}{2}(\kappa_2 - l^2)\, m''(x)\, h^2 \\ &\quad + \frac{l}{2}\left\{ \frac{f'(x)\,(\kappa_2^2 - \kappa_4)}{f(x)\,\kappa_2}\, m''(x) + \left(\kappa_2 - \frac{\kappa_4}{3\kappa_2} - \frac{2l^2}{3}\right) m'''(x) \right\} h^3 \\ &\quad + \frac{1}{2}\left(\left[\left\{ \frac{f''(x)}{2f(x)} - \left(\frac{f'(x)}{f(x)}\right)^2 \right\} (\kappa_4 - \kappa_2^2) - \frac{l f''(x)\,\kappa_5}{2f(x)\,\kappa_2} \right.\right. \\ &\quad \left. + \left\{ \frac{f''(x)}{f(x)} - \left(\frac{f'(x)}{f(x)}\right)^2 \right\} \frac{l^2\,(\kappa_2^2 - \kappa_4)}{\kappa_2} \right] m''(x) \\ &\quad + \frac{l f'(x)\,\{3l\,(\kappa_2^2 - \kappa_4) - \kappa_5\}}{3f(x)\kappa_2}\, m'''(x) \\ &\quad \left. + \frac{1}{2}\left\{ \frac{\kappa_4}{6} - \frac{l\kappa_5}{6\kappa_2} + \frac{l^2(3\kappa_2^2 - 2\kappa_4)}{3\kappa_2} - \frac{l^4}{2} \right\} m^{(iv)}(x) \right) h^4 + O_p(\delta_4)\,. \end{aligned} \tag{2.7}$$

Considering versions of this formula in the cases $l = 0, l_1, l_2$, and combining them to produce a formula for conditional bias for $\widetilde{m}(x)$ defined at (2.2), we see that the terms in h^2 and h^3 vanish if and only if

$$\begin{aligned} &\lambda_1(\kappa_2 - l_1^2) + \kappa_2 + \lambda_2(\kappa_2 - l_2^2) = 0\,, \\ &\lambda_1 l_1 + \lambda_2 l_2 = 0\,, \quad \lambda_1 l_1^3 + \lambda_2 l_2^3 = 0\,. \end{aligned}$$

Assuming only that $\lambda_1, \lambda_2 > 0$ and $l_1, l_2 \neq 0$, the latter two equations imply that $\lambda_1 = \lambda_2 = \lambda$ and $l_1 = -l_2 = l$, say. The first equation then gives $l = l(\lambda)$, where $l(\lambda)$ is defined by (2.3). Finally, the claimed bias expansion in Theorem 2.1 follows directly from (2.7).

Asymptotic properties of the conditional variance of $\widetilde{m}$ can be derived similarly, and are given in the next theorem.

Theorem 2.2 *Assume the conditions imposed on K and h in Theorem 2.1, that f has a bounded derivative in a neighbourhood of x, and that $v(u) = \mathrm{var}\,(Y|X = u)$ is bounded and continuous there. Assume that $\lambda_1 = \lambda_2 = \lambda > 0$ and $l_1 = -l_2 = l(\lambda)$. Then,*

$$\mathrm{var}\,\{\widetilde{m}(x)|X_1, \ldots, X_n\} = \frac{1}{nh}\,\frac{v(x)}{f(x)}\,V(\lambda) + o_p\{(nh)^{-1}\}\,,$$

where

$$\begin{aligned} V(\lambda) &= (2\lambda+1)^{-2}\Big[\big(2\lambda^2+1\big)\int K(u)^2\,du \\ &\quad +\big(6\lambda+1\big)\int K(u-l)\,K(u)\,du \\ &\quad +\tfrac{1}{2}\big(4\lambda+1\big)^2\int K(u-l)\,K(u+l)\,du \\ &\quad +\lambda(2\lambda+1)\kappa_2^{-1}\int u^2\big\{K(u)^2-K(u-l)\,K(u+l)\big\}\,du\Big]\,. \end{aligned} \tag{2.8}$$

Proof of Theorem 2.2. Since $\lambda_1=\lambda_2=\lambda>0$ and $l_1=-l_2=l(\lambda)$, the estimator at (2.2) may be expressed as

$$\widetilde{m}(x)=\big(2\lambda+1\big)^{-1}\big\{\lambda\,\widehat{m}(x|x+lh)+\widehat{m}(x|x)+\widehat{m}(x|x-lh)\big\}\,. \tag{2.9}$$

Denote the conditional variance of $\widehat{m}(u|x)$ by $\hat{\eta}(u|x)$, and the conditional covariance of $\widehat{m}(u|x)$ and $\widehat{m}(u|y)$ by $\hat{c}(u|x,y)$. The conditional variance of the regression mean may be written as

$$\begin{aligned} \operatorname{var}\{\widetilde{m}(x)|\,X_1,\ldots,X_n\} = \big(2\lambda+1\big)^{-2}\big\{&\lambda^2\,\hat{\eta}(x|x+lh)+\hat{\eta}(x|x)+\lambda^2\,\hat{\eta}(x|x-lh) \\ &+2\lambda\,\hat{c}(x|x-lh,x)+2\lambda\,\hat{c}(x|x,x+lh) \\ &+2\lambda^2\,\hat{c}(x|x-lh,x+lh)\big\}\,. \end{aligned} \tag{2.10}$$

Up to first order, Theorem 2.2 follows from expansions for each term on the right-hand side of (2.10). For the sake of brevity, we shall only give details for the expansion of $\hat{\eta}(x|x+lh)$. Other terms may be expanded by following similar arguments as below. Using the definition of $\widehat{m}(u|x)$ at (2.1), we may express

$$\begin{aligned} \hat{\eta}(x|x+lh) &= E\left[\{\hat{a}(x+lh)-lh\,\hat{b}(x+lh)\}^2|\,X_1,\ldots,X_n\right] \\ &\qquad -\big[\,E\{\hat{a}(x+lh)-lh\,\hat{b}(x+lh)|\,X_1,\ldots,X_n\}\big]^2 \\ &= \operatorname{var}\{\hat{a}(x+lh)|\,X_1,\ldots,X_n\}+(lh)^2\operatorname{var}\{\hat{b}(x+lh)|\,X_1,\ldots,X_n\} \\ &\qquad -2\,lh\operatorname{cov}\{\hat{a}(x+lh),\hat{b}(x+lh)|\,X_1,\ldots,X_n\} \\ &= T_1+T_2-2\,T_3\,. \end{aligned}$$

Formula (2.6) in the proof of Theorem 2.1 gives the following identities:

$$s_0(x)=nh\,f(x)\big\{1+o_p(1)\big\}\,,\quad s_1(x)=nh^3\,f'(x)\,\kappa_2\big\{1+o_p(1)\big\}\,,$$

$$s_2(x)=nh^3\,f(x)\,\kappa_2\big\{1+o_p(1)\big\}\,, \tag{2.11}$$

whence it follows that

$$s_0(x)\, s_2(x) - s_1(x)^2 = n^2 h^4 f(x)^2 \kappa_2 \{1 + o_p(1)\}. \tag{2.12}$$

From the definitions of $\hat{a}$ and $\hat{b}$ together with (2.11), we deduce that

$$\begin{aligned}
&\sum_{i=1}^{n} \{X_i - (x + l_1 h)\}^{t_1} \{X_i - (x + l_2 h)\}^{t_2} \\
&\qquad \times K\Big\{\frac{X_i - (x + l_1 h)}{h}\Big\} K\Big\{\frac{X_i - (x + l_2 h)}{h}\Big\} v(X_i) \\
&= nh^{t_1+t_2+1} f(x) v(x) \Big\{ \int (u - l_1)^{t_1} (u - l_2)^{t_2} K(u - l_1) K(u - l_2)\, du \Big\} \{1 + o_p(1)\},
\end{aligned}$$

for non-negative integers t_1, t_2. Using the above identity and (2.12), together with some cumbersome algebra, we see that

$$\begin{aligned}
T_1 &= \{n^2 h^4 f(x + lh)^2 \kappa_2\}^{-2} \big[s_2(x + lh)^2 \operatorname{var}\{r_0(x + lh) | X_1, \ldots, X_n\} \\
&\qquad + s_1(x + lh)^2 \operatorname{var}\{r_1(x + lh) | X_1, \ldots, X_n\} \\
&\qquad - 2\, s_1(x + lh)\, s_2(x + lh) \operatorname{cov}\{r_0(x + lh), r_1(x + lh) | X_1, \ldots, X_n\}\big] \\
&\qquad \times \{1 + o_p(1)\} \\
&= (nh)^{-1} f(x)^{-1} v(x) \Big\{ \int K(u)^2\, du \Big\} \{1 + o_p(1)\}, \\
T_2 &= \{n^2 h^3 l f(x + lh)^2 \kappa_2\}^{-2} \big[s_0(x + lh)^2 \operatorname{var}\{r_1(x + lh) | X_1, \ldots, X_n\} \\
&\qquad + s_1(x + lh)^2 \operatorname{var}\{r_0(x + lh) | X_1, \ldots, X_n\} \\
&\qquad - 2\, s_0(x + lh)\, s_1(x + lh) \operatorname{cov}\{r_0(x + lh), r_1(x + lh) | X_1, \ldots, X_n\}\big] \\
&\qquad \times \{1 + o_p(1)\} \\
&= (nh)^{-1} (l\kappa_2^{-1})^2 f(x)^{-1} v(x) \Big\{ \int u^2 K(u)^2\, du \Big\} \{1 + o_p(1)\}, \\
T_3 &= (nh)^{-1} l \kappa_2^{-1} f(x)^{-1} v(x) \Big\{ \int (u + l) K(u + l)^2\, du \Big\} \{1 + o_p(1)\} \\
&= o_p\{(nh)^{-1}\}.
\end{aligned}$$

The term of size $(nh)^{-1}$ in T_3 vanishes since, for a symmetric kernel K, $\int (u + l) K(u + l)^2\, du = 0$. Combining these results, we may prove that

$$\hat{\eta}(x | x + lh) = (nh)^{-1} f(x)^{-1} v(x) \Big\{ \int K^2 + (l\kappa_2^{-1})^2 \int u^2 K(u)^2\, du \Big\} \{1 + o_p(1)\}.$$

Other terms on the right-hand side of (2.10) may be expanded similarly, and we only state the results here:

$$\hat{\eta}(x | x) = (nh)^{-1} f(x)^{-1} v(x) \Big(\int K(u)^2\, du \Big) \{1 + o_p(1)\},$$

$$\begin{aligned}
\hat{\eta}(x|x \pm lh) &= (nh)^{-1} f(x)^{-1} v(x) \Big\{ \int K(u)^2 \, du \\
&\qquad + \big(l\kappa_2^{-1}\big)^2 \int u^2 \, K(u)^2 \, du \Big\} \big\{1 + o_p(1)\big\}, \\
\hat{c}(x|x \pm lh, x) &= (nh)^{-1} f(x)^{-1} v(x) \Big\{ \int K(u) \, K(u \pm l) \, du \\
&\qquad \pm \big(l\kappa_2^{-1}\big) \int (u \pm l) \, K(u \pm l) \, K(u) \, du \Big\} \big\{1 + o_p(1)\big\}, \\
\hat{c}(x|x - lh, x + lh) &= (nh)^{-1} f(x)^{-1} v(x) \Big\{ \kappa_2^{-1}(\kappa_2 + 2l^2) \int K(u + l) \, K(u - l) \, du \\
&\qquad - (l\kappa_2)^2 \int (u^2 - l^2) \, K(u + l) \, K(u - l) \, du \Big\} \big\{1 + o_p(1)\big\}.
\end{aligned}$$

Substituting these formulae into (2.10) gives

$$\begin{aligned}
\operatorname{var}\{\widetilde{m}(x) | \, X_1, \ldots, X_n\} &= (nh)^{-1} (2\lambda + 1)^{-2} f(x)^{-1} v(x) \Big[(2\lambda^2 + 1) \int K(u)^2 \, du \\
&\qquad + 2\lambda\kappa_2^{-1} \big(2\kappa_2 + l^2\big) \int K(u - l) \, K(u) \, du \\
&\qquad + 2\big\{\lambda\kappa_2^{-1}(\kappa_2 + l^2)\big\}^2 \int K(u - l) \, K(u + l) \, du \\
&\qquad + 2\big(\lambda l \kappa_2^{-1}\big)^2 \int u^2 \big\{K(u)^2 - K(u - l) \, K(u + l)\big\} \, du \Big] \\
&\qquad \times \big\{1 + o_p(1)\big\}.
\end{aligned}$$

Putting $l(\lambda) = \{(1 + 2\lambda)\kappa_2/(2\lambda)\}^{1/2}$ yields the desired result.

There are of course versions of both theorems for kernels that are not compactly supported, although the regularity conditions depend to some extent on the rate of decay of the tails of the kernel. In the case of Normal kernel, defined by $K(u) = (2\pi)^{-1/2} \exp(-u^2/2)$, it is sufficient to impose the following additional conditions: in both theorems, assume that f is bounded on the real line $\mathbb{R}$, and that $h = o\{(\log n)^{-1/2}\}$; and in Theorem 2.1 (respectively, Theorem 2.2), assume that m (respectively, v) is bounded on $\mathbb{R}$.

Taking $\lambda = \infty$ in the definition of $\widetilde{m}$, we obtain $\widetilde{m}$ as a linear combination of $\widehat{m}(x|x + \kappa_2^{1/2} h)$ and $\widehat{m}(x|x - \kappa_2^{1/2} h)$, denoted by $\widetilde{m}_\infty(x)$. This skewed estimator generally has larger variance than $\widetilde{m}$ for finite λ, as we shall see below. Its bias,

$$\{8f(x)\}^{-1}(\kappa_2^2 - \kappa_4) \big\{f''(x) \, m''(x) + 4f'(x) \, m'''(x) + f(x) \, m^{(iv)}(x)\big\},$$

can be either greater than or less than that of $\widetilde{m}$ for finite λ, depending on the shape of m and f.

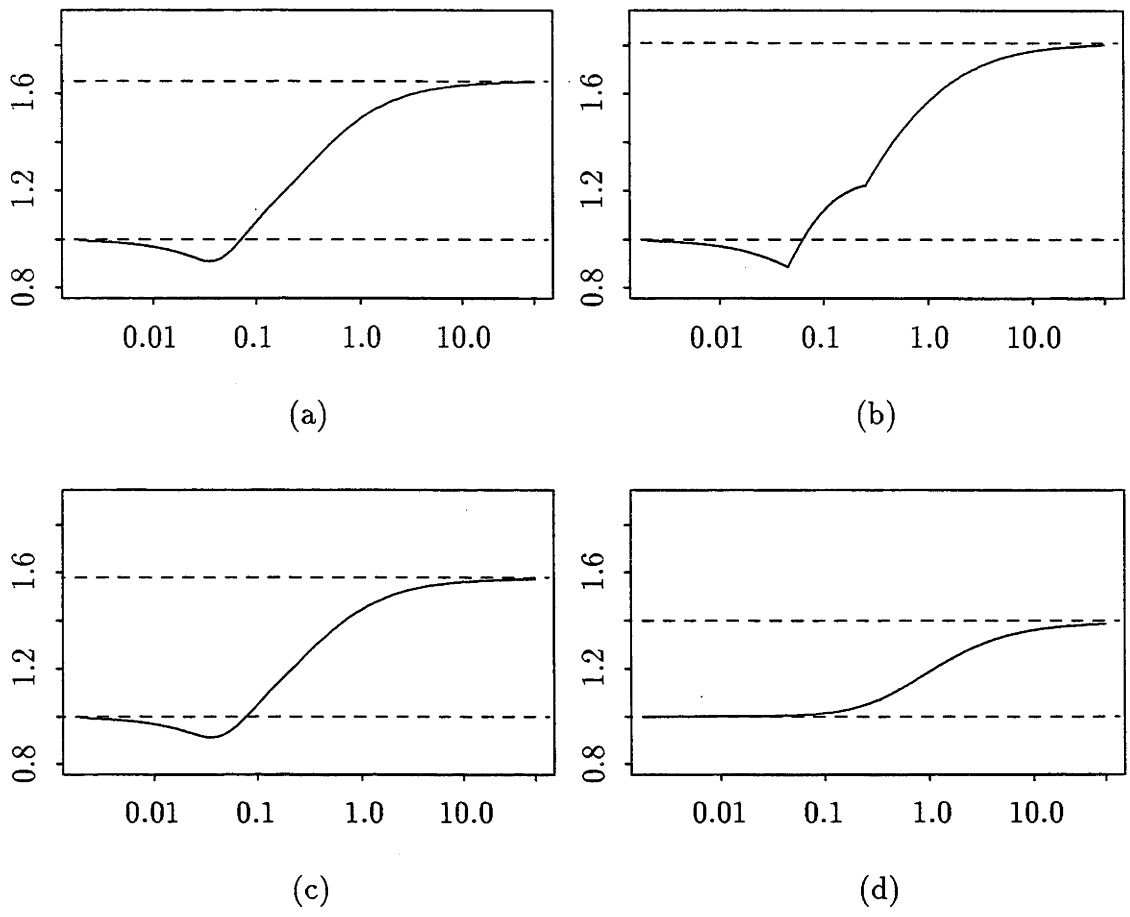


Figure 2.2: Graphs of the function $V(\lambda)/V(0)$ against λ (on logarithmic scales). Panels (a) to (d) depict the Epanechnikov kernel, the Uniform kernel, the biweight kernel and the Normal kernel, respectively. The respective values of $V(\infty)/V(0)$ are $1.65, 1.81, 1.58$ and $\frac{1}{4}(3+7e^{-1}) = 1.39$.

It is clear that choice of λ can affect the variance of $\widetilde{m}$. The asymptotes of the function V at $\lambda = 0$ and $\lambda = \infty$ can be deduced from (2.8): $V(0) = \int K^2$, and

$$\begin{aligned} V(\infty) &= \tfrac{1}{2}\int K^2 + 2\int K\big(u-\kappa_2^{1/2}\big)\,K\big(u+\kappa_2^{1/2}\big)\,du \\ &\quad + \tfrac{1}{2}\kappa_2^{-1}\int u^2\left\{K(u)^2 - K\big(u-\kappa_2^{1/2}\big)\,K\big(u+\kappa_2^{1/2}\big)\right\}du\,. \end{aligned}$$

Depending on the choice of K, V can have a minimum at a point λ satisfying $0 < \lambda < \infty$, or have its minimum at $\lambda = 0$. The former situation arises for the

Epanechnikov, Uniform and biweight kernels, where the respective minimising values of λ are 0.0352, 0.0455, 0.0352, and the respective values of $\{\min_\lambda V(\lambda)\}/V(0)$ are 0.908, 0.885 and 0.912. In the Normal case, V is a strictly increasing function, and so the minimum occurs at $\lambda = 0$. Figure 2.2 depicts graphs of $V(\lambda)/V(0)$ against λ for the four different kernels just mentioned.

2.4 Left- and Right-skewed Estimators

We saw in the last section that a linear combination of skewed estimators can be used to reduced bias by two orders of magnitude, to h^4 compared with order h^2 for local linear smoothers. An alternative approach, based on a single version of $\widehat{m}(u|x)$, may be used to estimate m with bias of order h^3. Its operation may be explained intuitively as follows. Observe from Figure 2.1 that the expected value of a line fitted to the curve above a point whose abscissa is $x + lh$ will cut the curve at some point, whose abscissa is u say, and so the line segment will have zero bias as an estimator of $m(u)$. If $l = l_0$ is chosen appropriately then we may take $u = x$. Such an "ideal" l necessarily depends on x through the unknown curve, but to first order it is independent of x: $l_0 = \pm\kappa_2^{1/2} + O(h)$, where either sign may be employed. Therefore, using $\widetilde{m}_+(x) = \widehat{m}(x|x + \kappa_2^{1/2}h)$ or $\widetilde{m}_-(x) = \widehat{m}(x|x - \kappa_2^{1/2}h)$ instead of $\widehat{m}(x) = \widehat{m}(x|x)$ to estimate $m(x)$ reduces bias by one order of magnitude, from $O(h^2)$ to $O(h^3)$. We call $\widetilde{m}_-$ a *left-skewed estimator* since the kernel weights are centred to the left of x, at which we wish to estimate the curve. Similarly, we call $\widetilde{m}_+$ a *right-skewed estimator.*

Variance of $\widetilde{m}_\pm$ is affected only by a constant factor compared with $\widehat{m}$, and not by an order of magnitude. The next theorem makes this explicitly clear.

Theorem 2.3 *Assume that K is nonnegative, bounded, symmetric and compactly supported, with $\int K = 1$; and that $h = h(n) \to 0$ and $nh \to \infty$. Then (a) if m has three bounded, continuous derivatives in a neighbourhood of x, if f has one bounded, continuous derivative there, and if $f(x) > 0$,*

$$E\{\widetilde{m}_\pm(x) - m(x)|X_1, \ldots, X_n\} = B_\pm(x)\, h^3 + o_p\{h^3 + (nh)^{-1/2}\}\,, \tag{2.13}$$

where $B_\pm(x) = \pm\frac{1}{2}\kappa_2^{-1/2}(\kappa_2^2 - \kappa_4)\,\{f'(x)m''(x)f(x)^{-1} + \frac{1}{3}m'''(x)\}$; and (b) if f has a bounded derivative in a neighbourhood of x, and v is bounded and continuous there,

$$\text{var}\,\{\widetilde{m}_\pm(x)|X_1, \ldots, X_n\} = (nh)^{-1}\, v(x)\, f(x)^{-1}\, V_1 + o_p\{(nh)^{-1}\}\,, \tag{2.14}$$

where $V_1 = \int K^2 + \kappa_2^{-1} \int u^2 K(u)^2 \, du$.

Proof of Theorem 2.3. The bias expansion is obtainable from (2.5) in the proof of Theorem 2.1. Using the notations in Theorem 2.1,

$$\begin{aligned}\hat{\mu}(x|x+lh) &= m(x) + \tfrac{1}{2}(\kappa_2 - l^2)\, m''(x)\, h^2 \\ &\quad + \frac{l}{2}\left\{ \frac{f'(x)\,(\kappa_2^2 - \kappa_4)}{f(x)\,\kappa_2}\, m''(x) + \left(\kappa_2 - \frac{\kappa_4}{3\kappa_2} - \frac{2l^2}{3}\right) m'''(x) \right\} h^3 \\ &\quad + o_p\{h^3 + (nh)^{-1}\}\,.\end{aligned}$$

Putting $l = \pm\kappa_2^{1/2}$ yields the desired bias expansion. The asymptotic variance follows directly from the proof of Theorem 2.2.

Note that $\widetilde{m}_+$ and $\widetilde{m}_-$ are special cases of $\widetilde{m}$ defined at (2.2), using a highly asymmetric choice of weights. The quantity V_1 introduced in Theorem 2.3 always exceeds $\int K^2$, and so the asymptotic variance of $\widetilde{m}_+$ and $\widetilde{m}_-$ always exceeds that of the standard local linear estimator $\widehat{m}$ (see (1.14)), although only by a constant factor, not an order of magnitude. Indeed, for the Epanechnikov, Uniform, biweight and Normal kernels, the respective values of $V_1 / \int K^2$ are $\frac{12}{7}$, $\frac{3}{2}$, $\frac{18}{11}$ and $\frac{3}{2}$. As discussed in the previous section, a linear combination of $\widetilde{m}_+$ and $\widetilde{m}_-$ further reduces bias to size h^4.

2.5 Further Issues in Skewing

Of course, there are other versions of skewed estimators which can improve on the bias of the classical linear estimator. One such estimator takes the form

$$\lambda\, \widehat{m}(x|x) + (1-\lambda)\, \widehat{m}(x|x+lh)\,,$$

where $0 < \lambda < 1$. It may be shown that the choice of $l = \pm\{\kappa_2/(1-\lambda)\}^{1/2}$ reduces bias to order h^3. Variance remains of size $(nh)^{-1}$, although it is generally larger than $\int K^2$.

The expansions of bias and variance given in Theorems 2.1 and 2.2 are valid only in a conditional sense, and cannot be generalised to unconditional expansions without adjusting the estimators. To appreciate why, observe that the denominators

in the definitions of $\widehat{m}(x)$ and $\widehat{m}(u|x)$ can take the value zero with positive probability. Indeed this will always happen if there is just one design point in the interval $\mathcal{I}_x = (x - ch, x + ch)$, where c is chosen so that the support of K is the interval $(-c, c)$. Here the definition of $\widehat{m}(u|x)$ generally has the form "non-zero number divided by 0". When there are no design points in $\mathcal{I}_x$ the ratio is $0/0$. In less extreme cases where there are two or more points in $\mathcal{I}_x$, the denominator is non-zero but the estimator can nevertheless fluctuate erratically.

Several methods for overcoming these numerical difficulties have been discussed in Section 1.5 in the case of $\widehat{m}(x)$. They include incorporating a ridge parameter into the denominator (Fan, 1993; Seifert and Gasser, 1996a, 1996b) or imputing new design points in places where the original design sequence is sparse (Hall and Turlach, 1997b); and they render bias and variance formulae for standard local linear smoothing estimators valid in an unconditional sense. The procedures of Seifert and Gasser (1996a, 1996b) and Hall and Turlach (1997b) are also appropriate for our estimators, and allow the bias and variance expansions in Theorems 2.1–2.3 to be stated in an unconditional sense.

In the case of the ridge parameter method, the range of allowable ridges is slightly narrower than for standard local linear smoothing, since bias is of smaller order there and we must ensure that the ridge does not interfere with bias (see Hall and Marron, 1997). It is sufficient that the ridge be of size $n^{-\alpha}\{(nh^2)^2 + nh\}$ for some $\alpha > 0$. For example, if we redefine $\widehat{m}(u|x)$ by replacing the denominator $s_0(s)\,s_2(s) - s_1(x)^2$ by $s_0(x)\,s_2(x) - s_1(x)^2 + r(n,h)$, where $r(n,h)$ is a non-negative sequence satisfying $r(n,h) \sim Cn^{-\alpha}\{(nh^2)^2 + nh\}$ for positive constants C and α; and if the conditions of Theorems 2.1 and 2.2 are valid; then the bias and variance expansions there are valid unconditionally:

$$\begin{aligned} E\{\widetilde{m}(x) - m(x)\} &= B(x)\,h^4 + o\{h^4 + (nh)^{-1/2}\}\,, \\ \text{var}\,\{\widetilde{m}(x)\} &= (nh)^{-1}v(x)f(x)^{-1}V(x) + o\{h^8 + (nh)^{-1}\}\,. \end{aligned}$$

Techniques of Hall and Turlach (1997b) are applicable without change, and also produce these unconditional expansions. They will be explored further in the numerical studies in the next section. Entirely analogous remarks apply to the estimators $\widetilde{m}_\pm$ and $\frac{1}{2}(\widetilde{m}_+ + \widetilde{m}_-)$.

Unlike $\widehat{m}(x)$, the estimators $\widetilde{m}(x)$ and $\widetilde{m}_\pm(x)$ suffer from edge effects at the ends of the design interval $\mathcal{I}$. These reduce performance to that of $\widehat{m}(x)$, in terms of order of magnitude of bias, at the very ends of $\mathcal{I}$. It is worth stressing that, for

arbitrary fixed l, there is no component of size h in the bias of $\widehat{m}(x|x+lh)$. This is clear from equation (2.7), under the assumptions that led to that result, although it may be seen more generally from the definition of $\widehat{m}(u|x)$ at (2.1). Variance remains of order $(nh)^{-1}$, uniformly in the interval of estimation.

Therefore, if one wishes to estimate right to the ends of $\mathcal{I}$ at the same order of bias, it is necessary to correct for edge effects. The edge problems of $\widetilde{m}$ may be overcome in standard ways, for example by employing $\widetilde{m}$ right up to the point where edge effects start to occur, and fitting locally a cubic polynomial from there to the boundary. The estimator $\widetilde{m}_+$, however, does not suffer edge effects at the left-hand end of $\mathcal{I}$, and, likewise, $\widetilde{m}_-$ does not have problems at the right-hand end.

Standard methods for bandwidth selection that require only linearity in the responses Y_i are applicable without modification for our skewed estimators. Such methods include cross-validation or generalised cross-validation (e.g. Fan and Gijbels, 1996, p. 149ff) or the m-out-of-n bootstrap (see Faraway and Jhun, 1990, for the density estimation case). Plug-in rules (e.g. Fan and Gijbels, 1996, p. 152ff) are also appropriate, but since the bias formulae in Theorems 2.1 and 2.3 are more complex than in the case of local linear regression, they are less attractive. However, this is a problem for all high-order methods since high-order derivatives, such as those of the design density, are difficult to estimate unless data are plentiful. In those cases, simpler approaches such as cross-validation are suggested.

2.6 Numerical Performance

Our numerical studies, which assess the finite-sample performance of the skewed estimators, have two parts. First, we compare the classic local linear estimator with our skewed estimators. Secondly, we compare performance of the fourth-order version of our methods with ridged local cubic regression.

2.6.1 Comparison with Local Linear Estimators

We conducted a simulation study to compare the finite-sample performance of the local linear method with that of various versions of our skewed estimators, including $\widetilde{m}_\pm(x) = \widehat{m}(x \pm \kappa_2^{1/2} h)$,

$$\widetilde{m}_\lambda(x) = (2\lambda+1)^{-1}\left\{\lambda\,\widehat{m}(x|x-lh) + \widehat{m}(x|x) + \lambda\,\widehat{m}(x|x+lh)\right\}, \qquad (2.15)$$

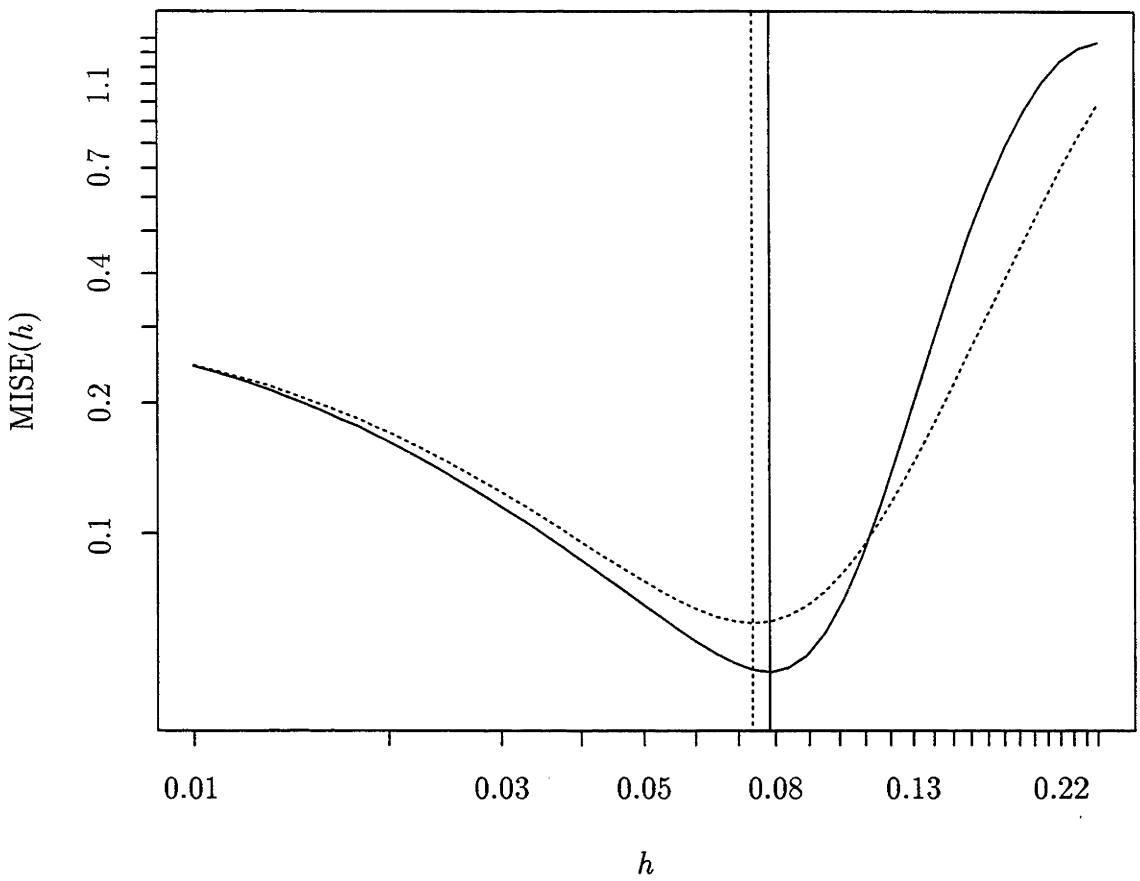


Figure 2.3: Comparison between MISE performance of the estimator at (2.16), and that of the local linear fit, with $n = 100$ *and* $m = m_2$. The curves depict MISE as a function of bandwidth h, with the solid and dotted lines representing the estimator $\widetilde{m}_\lambda$ and the local linear fit respectively.

with $l = \{(1+2\lambda)\,\kappa_2/(2\lambda)\}^{1/2}$, and $\widetilde{m}_\infty = \frac{1}{2}(\widetilde{m}_+ + \widetilde{m}_-)$. The bias-reduction properties of these estimators are of course clearest in the case of regression means that are significantly nonlinear, since even the general form of the local linear estimator $\widehat{m}(u|x)$ has exactly zero bias for a linear target. We took as our target the sine function on the interval $\mathcal{I} = [0,1]$ with $k = 1, 2$ or 3 wavelengths, i.e. $m(x) = m_k(x) = 2\sin(2k\pi x)$, for Uniformly distributed design points X_i, Normal N(0,0.5) errors, and sample sizes $n = 50, 100, 200$ and 400. Only the two-wave mean, $m = m_2$, will be discussed in detail, although numerical results for other values of k are given in Tables 2.1 and 2.2.

Number of waves (k)	Sample size n 50	100	200	400
1	1.30	1.38	1.46	1.57
2	1.22	1.30	1.38	1.47
3	1.17	1.25	1.33	1.42

Table 2.1: *Ratio of minimum MISE values of $\widehat{m}$ and $\widetilde{m}_\lambda$.* For each sample size, n, and mean function, m_k, the ratio ρ, of the minimum MISE of the local linear smoother $\widehat{m}$ to that of $\widetilde{m}_\lambda$, defined at (2.16), is tabulated.

We used a grid of bandwidths that consisted of 51 logarithmically equally-spaced points in the interval $[0.01, 0.25]$. All mean integrated squared error (MISE) curves were computed as averages of 1000 replications of integrated squared error (ISE) curves. Each ISE curve was in turn obtained by, for each h in the grid, evaluating pointwise squared errors at 401 equally-spaced points in $\mathcal{I}$. The trapezoidal rule was employed to calculate integrated squared error.

We used the Epanechnikov kernel, $K(u) = \frac{3}{4}(1-u^2)\,I(|\,u\,| \leq 1)$, throughout. In this setting, the minimum variance of $\widetilde{m}_\lambda$ is attained at $\lambda = 0.0352$ (see Section 2.3), in which case $\widetilde{m}_\lambda$ becomes

$$\widetilde{m}_\lambda(x) = 0.0329\,\widehat{m}(x|x-1.74h) + 0.9342\,\widehat{m}(x|x) + 0.0329\,\widehat{m}(x|x+1.74h)\,. \quad (2.16)$$

A simple linear interpolation rule (Hall and Turlach, 1997b) was employed to guard against sparse design and ensure at least three design points in each interval $(x-h, x+h)$ for $x \in \mathcal{I}$. Extra data were generated outside $\mathcal{I}$ to avoid boundary problems.

The solid curve in Figure 2.3 represents the MISE of $\widetilde{m}_\lambda$, given by (2.16), as a function of bandwidth h, in the case $m = m_2$ and $n = 100$. The dotted curve represents MISE of the local linear fit. Vertical lines are drawn through their respective minimisers. As expected, the overall minimum is lower for $\widetilde{m}_\lambda$ than for the local linear estimator $\widehat{m}(x)$, and $\widetilde{m}_\lambda$ performs better for very small bandwidths owing to its smaller variance component. But $\widehat{m}$ is superior for relatively large bandwidths, reflecting inadequacy of the asymptotic bias formula in Theorem 2.1 in this case. Relative "efficiency", ρ, of the two estimators, represented as the ratio of minimum MISE for the local linear estimator to that for $\widetilde{m}_\lambda$, equals 1.32. Values of ρ for the two other regression means, m_1 and m_3, and for other sample sizes, are given

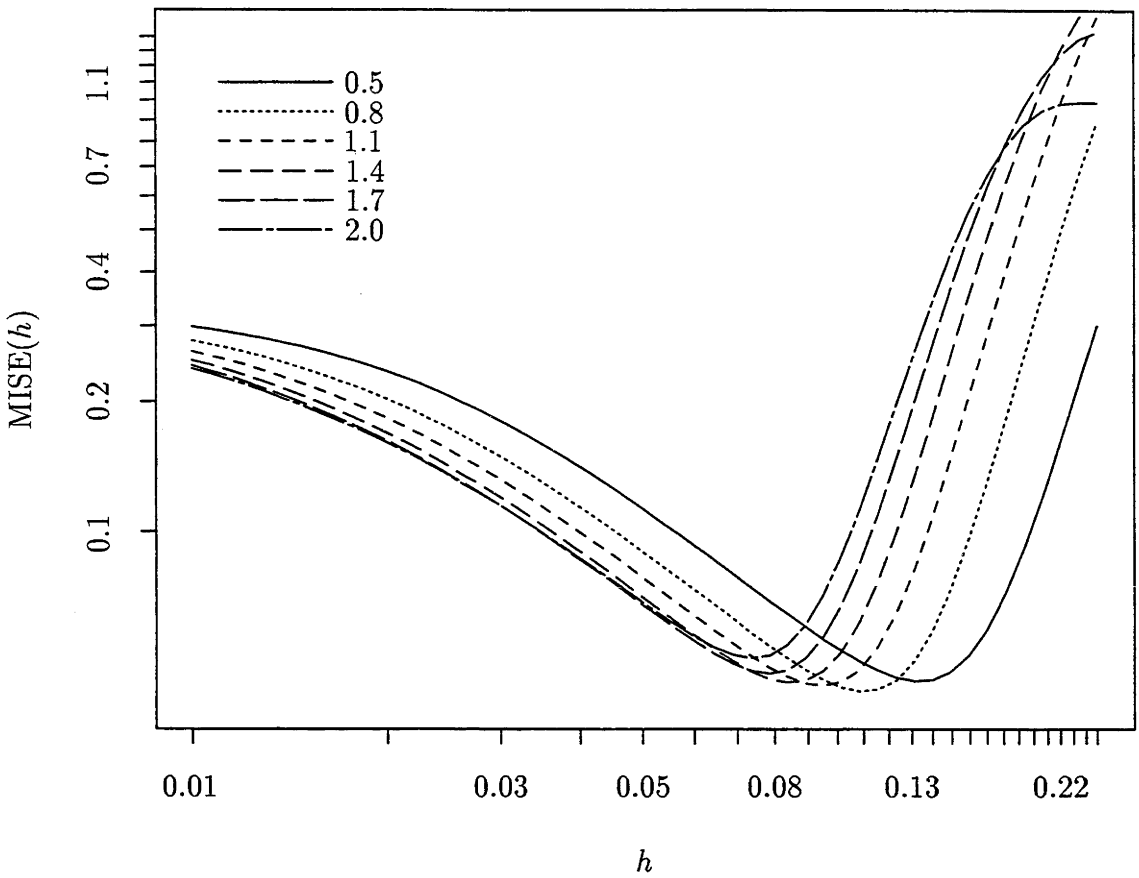


Figure 2.4: Comparison of MISE performance of the estimator at (2.15) for different values of l, with $n = 100$ and $m = m_2$. The legends on the graphs indicate the values of l used to construct the estimator at (2.15).

in Table 2.1.

To study the effect of choice of λ on the performance of the estimator at (2.15), we varied l in the range $[0.5, 2.0]$, which corresponds to adjusting $\lambda = \kappa_2/\{2(l^2-\kappa_2)\}$ from 0.0263 to 2.0, with $\kappa_2 = 0.2$. The results in the case $n = 100$ and $m = m_2$ are shown in Figure 2.4. The effect is largely to translate the MISE curve along the h axis, preserving both its shape and its height. Thus, the minimum value attained by MISE is relatively immune to fluctuations in choice of λ over a moderately wide range, despite the rather steep increase in variance for large λ's, depicted in panel (a) of Figure 2.2. The value of the bandwidth at which the minimum occurs, however, shows a marked tendency to decrease with increasing λ.

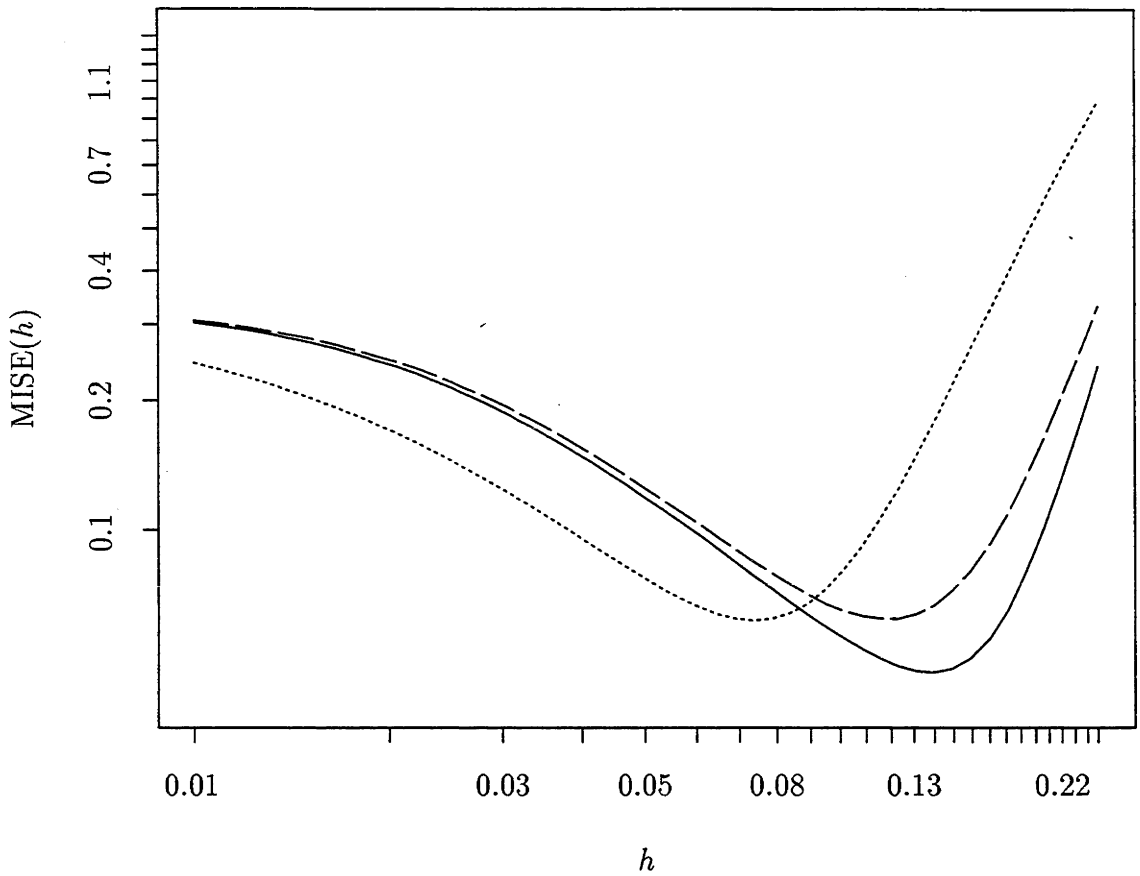


Figure 2.5: MISE curves for estimators $\widetilde{m}_{\pm}$ and $\widetilde{m}_{\infty}$, with $n = 100$ and $m = m_2$. The long-dashed, unbroken and dotted lines represent MISE curves for $\widetilde{m}_{\pm}$, $\widetilde{m}_{\infty}$ and the local linear fit, respectively.

Figure 2.5 compares MISE performance of $\widetilde{m}_{\pm}$, represented by the long-dashed line, of $\widetilde{m}_{\infty}$, producing the unbroken line, and of the classical local linear smoother, given by the dotted line. For the sample size and regression mean used in this study, i.e. $n = 100$ and $m = m_2$, the estimators $\widetilde{m}_{\pm}$ actually perform slightly less well than the local linear smoother, in that the ratio of the minimum mean integrated squared error of the latter to that of the former is $\rho = 0.989$. Performance of $\widetilde{m}_{\pm}$ does improve for larger sample sizes, however; for example, when $n = 400$ and $k = 1, 2, 3$, the value of ρ increases to $1.21, 1.13$ and 1.10, respectively. On the other hand, the MISE performance of $\widetilde{m}_{\infty}$, demonstrated in Table 2.2, consistently improves on that of both $\widetilde{m}_{\pm}$ and local linear smoothing over a wide range, as suggested by Figure 2.5.

Number of waves (k)	Sample size n 50	100	200	400
1	1.31	1.40	1.50	1.61
2	1.26	1.32	1.40	1.50
3	1.21	1.27	1.35	1.44

Table 2.2: *Ratio of minimum MISE values of $\widehat{m}$ and $\widetilde{m}_\infty$.* For each sample size, n, and mean function, m_k, the ratio ρ, of the minimum MISE of the local linear smoother $\widehat{m}$ to that of $\widetilde{m}_\infty$, is tabulated.

For the present choice of target, $\widetilde{m}_\infty$ marginally outperforms $\widetilde{m}_\lambda$ (compare Tables 2.1 and 2.2).

2.6.2 Comparison with Local Cubic Estimators

In this section, we look further into our skewing methods and compare them with higher-order local polynomial methods. We deliberately do not explore the estimators $\widetilde{m}_\pm$ here, since they tend to estimate a symmetric peak consistently to one side or other of its actual locations, depending on whether they are left- or right-skewed estimators. This asymmetry is rather undesirable, especially in more exploratory curve estimation problems. Thus, we refrain from investigating $\widetilde{m}_\pm$ extensively.

As already discussed in Section 1.5, several techniques for overcoming numerical problems with sparse design have been suggested in the case of local linear methods. In principle, similar techniques may be employed in the case of high-order polynomials, but they are awkward to apply, not least because they involve inversion of a $(p+1) \times (p+1)$ matrix if we are fitting a polynomial of degree p. This increase in dimensionality leads to difficulties from at least two sources: first, through greater likelihood of encountering sparse design problems (an aspect of the "curse of dimensionality"), evident in increased tendency for the matrix to be singular or nearly singular; and, secondly, through increased difficulty implementing corrections such as those based on ridging, shrinkage and imputation. By way of comparison, since skewing permits us to achieve the performance of a $p = 2$ or 3 method while using a $p = 1$ approach, it allows sparse design to be overcome relatively easily.

The simulation study in this section mainly compared the finite-sample perfor-

mance of local cubic estimator, $\widehat{m}_c$, with that of the minimum variance version $\widetilde{m}_\lambda$ (defined at (2.15)) and $\widetilde{m}_\infty$. We took the target, m, to be a combination of two sine functions,

$$m(x) = m_k(x) = \tfrac{2}{5}\{3\sin(2k\pi x) + 2\sin(3\pi x)\},$$

on the interval $\mathcal{I} = [0,1]$ for $k = 1, 2$ or 3. For the sake of brevity, we present our results only in the case where $m = m_2$, errors are Normal $\mathrm{N}(0, 0.5^2)$, and sample size is $n = 50$. The variance of the errors here is slightly smaller compared with the previous simulation study, since the range of the regression curve is relatively narrower than that of the sine function there. Our design density has distribution $0.6\,B(8,4) + 0.4\,U(\mathcal{I})$, where $B(p,q)$ and $U(\mathcal{I})$ denote respectively the Beta distribution with parameters p and q and the Uniform distribution on the interval $\mathcal{I} = [0,1]$. Similar results were obtained with other targets (e.g. $m(x) = \sin(2k\pi x) + \frac{9}{4}(2x-1)^2$), sample sizes ($n = 100, 200$ and 400), and other choices of design (e.g. $0.5\,B(4,4) + 0.5\,U(\mathcal{I})$). As before we used the Epanechnikov kernel throughout, and $\widetilde{m}_\lambda$ is given by (2.16). The MISE curves were obtained in exactly the same way, except that the grid of bandwidths now consisted of 51 logarithmically equally-spaced points in the interval $[0.01, 0.35]$.

When using skewing methods, the $\nu = 3$ linear interpolation rule of Hall and Turlach (1997b) was employed to guard against sparse design and ensure at least three design points in each interval $(x-h, x+h)$, for $x \in \mathcal{I}$. Neither this approach nor the shrinkage technique has a straightforward analogue in the case of local cubic smoothing. This is not difficult to see: if the interval $(x-h, x+h)$ consists only of pseudo data points, the matrix $\mathbf{X}^T\mathbf{W}\mathbf{X}$ may still be singular. Therefore, we used ridge methods there, employing a scale multiple of the 4×4 identity matrix, $\epsilon\mathbf{I}$, as the ridge. Except for the addition of this ridge to the matrix

$$\begin{pmatrix} S_0 & S_1 & S_2 & S_3 \\ S_1 & S_2 & S_3 & S_4 \\ S_2 & S_3 & S_4 & S_5 \\ S_3 & S_4 & S_5 & S_6 \end{pmatrix}$$

with $S_j = \sum_{i=1}^n (X_i - x)^j K_h(X_i - x)$ for $j = 0, 1, \ldots, 6$, our local polynomial methods were identical to those described by Ruppert and Wand (1994) and Fan and Gijbels (1996, p.58ff). See also Section 1.3. Extra data were generated outside $\mathcal{I}$ to avoid boundary problems. For our choice of design density, this could be done

using a Uniform distribution there, provided both the Beta parameters in the design density were taken to be greater than or equal to 3.

As expected from the theory, skewing is closely comparable with local cubic smoothing in MISE terms, when the ridge parameter is chosen optimally. From this viewpoint, the choice of parameters for our simulation study slightly favours local cubic methods. However, in computational or numerical terms, the skewing approach is noticeably superior, for several reasons. First, each calculation of local cubic smoothers requires the solution of a four-parameter optimisation problem, involving inversion of a 4×4 matrix. The calculation of the skewed estimator, however, demands only three solutions to two-parameter problems, each of which is solvable without matrix inversion. This leads to significant computational savings, as well as greater numerical stability.

Secondly, while the skewed estimator demands only choice of the bandwidth parameter, local cubic smoothing requires selection of both the bandwidth and the ridge. Although in theory the ridge only smooths high-order terms in expansions of mean squared error, in practice it plays a significant role as a smoothing parameter. The problems experienced by ridged local cubic methods are more severe than ridged local linear methods, because of the significantly greater sensitivity of the former to data sparseness, and hence of the greater reliance of local cubic methods on the ridge. The selection of the ridge, however, can be entirely avoided in local linear smoothing using techniques such as those of Hall and Turlach (1997b).

Indeed, for small to moderate sample sizes, the ridge and the bandwidth interact with one another, and have to be selected jointly by applying a bivariate smoothing-parameter choice algorithm such as bivariate cross-validation. The strong interaction is illustrated in panel (a) of Figure 2.6, which depicts MISE as a function of bandwidth for various choices of the ridge parameter, ϵ. Among the values considered there ($\epsilon = 10^{-2m}$ for $m = 0, 1, \ldots, 5$), $\epsilon = \epsilon_0 = 10^{-6}$ is optimal, but note that MISE increases steeply on either side of the optimal bandwidth. Therefore, errors in bandwidth choice, when using the optimal ridge, will be significant. The next smallest choice of the ridge suffers from this problem even more acutely. For the next largest and the next-but-one largest choice of ridge, the optimal bandwidths are only 84% and 42%, respectively, of their values in the case of ϵ_0, and the increases in minimal MISE can be substantial (they are 23% and 91% in these respective cases).

The MISE curves for two different skewed estimators ($\widetilde{m}_\infty, \widetilde{m}_\lambda$), the optimally

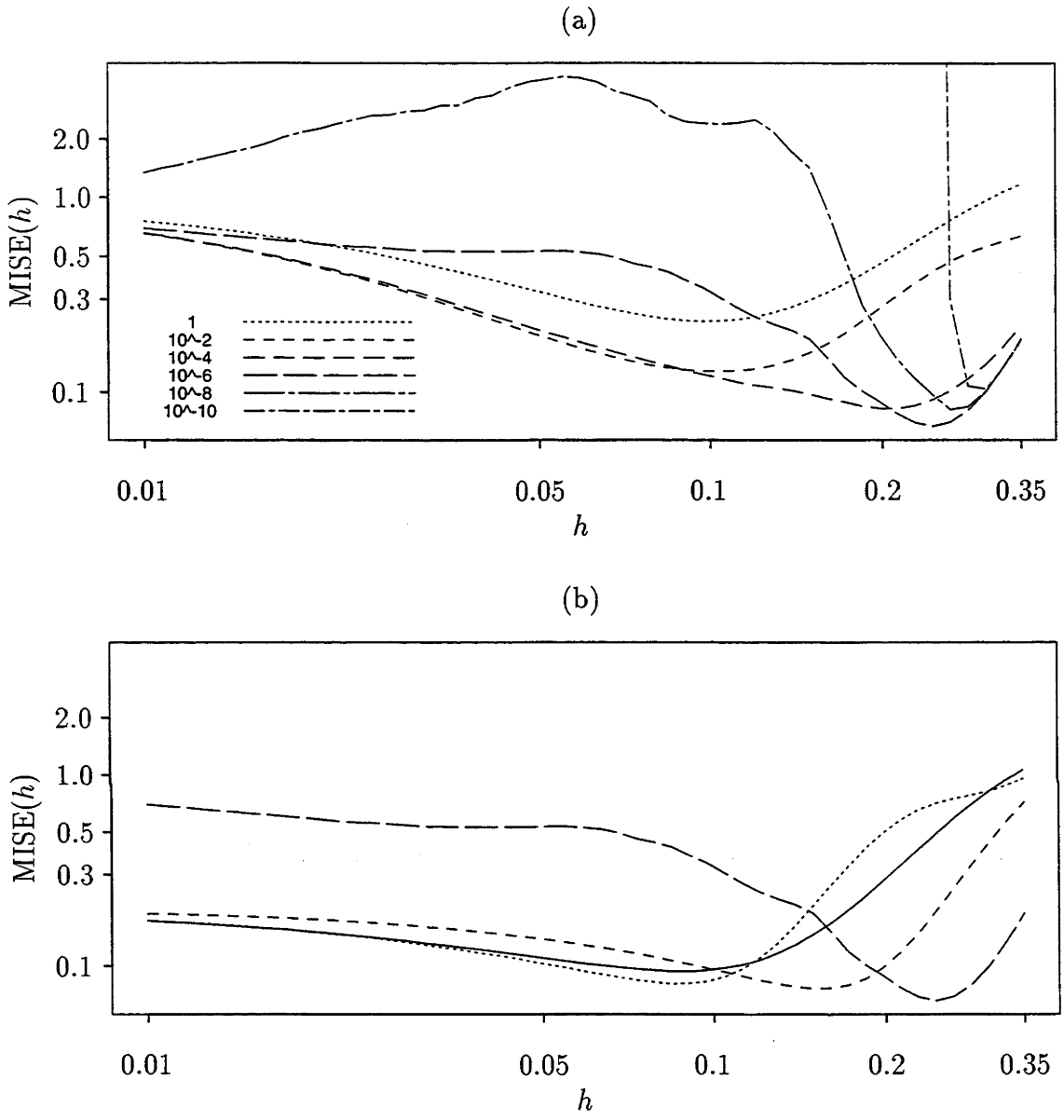


Figure 2.6: Mean integrated squared error comparison of skewing and ridged local cubic smoothing, with $n = 50$. Mean integrated squared error curves for various ridged versions of $\widehat{m}_c$ are displayed in panel (a). The legend in the graph indicates the values of ϵ used to construct $\widehat{m}_c$. Panel (b) depicts the curves for the "classic" local linear smoother, the skewed local linear smoothers $\widetilde{m}_\lambda$ and $\widetilde{m}_\infty$, and the optimal local cubic smoother $\widehat{m}_{c0}$, represented by solid, dotted, short-dashed and long-dashed lines respectively. The vertical axes in both panels have the same range and scale, and are logarithmically scaled.

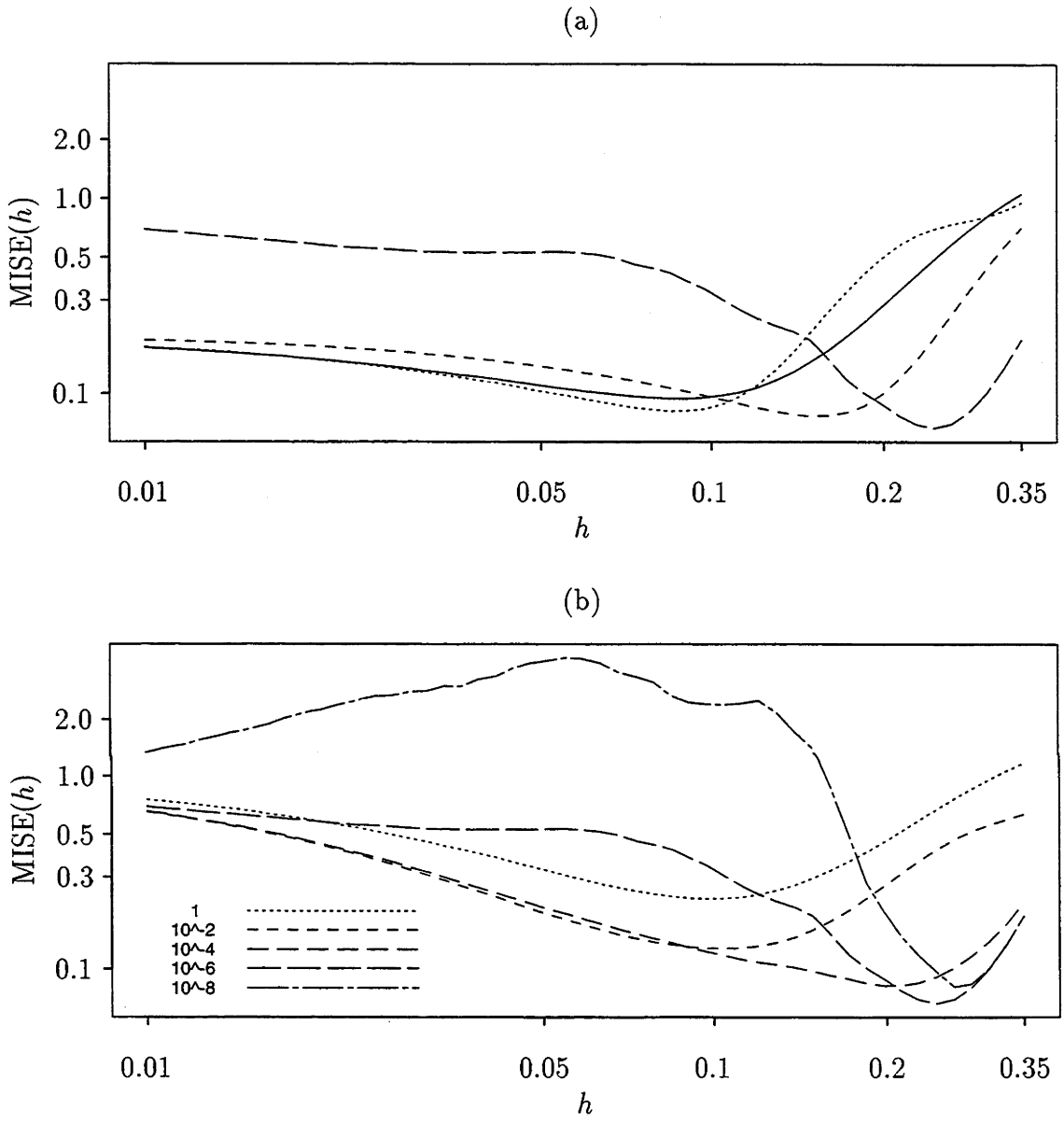


Figure 2.7: Mean squared error comparison of local performance of skewing and local cubic smoothing, with $n = 50$. Panels (a) and (b) show the mean squared error curves for different estimators at $x_1 = 0.1363$ and $x_2 = \frac{2}{3}$, respectively. Each panel illustrates curves for $\widetilde{m}_\lambda$, $\widetilde{m}_\infty$, $\widehat{m}_{c0}$ and $\widehat{m}_c$, with $\epsilon = 10^{-2}$, represented by dotted, short-dashed, long-dashed and dot-dashed lines respectively. The vertical axes are logarithmically scaled.

ridged local cubic estimator ($\widehat{m}_{c0}$, say, based on $\epsilon = \epsilon_0$) and, for the sake of completeness, the "classic", unskewed local linear estimator, are depicted in panel (b) in Figure 2.6. The MISE curve for the local cubic estimator is characterised by a sharp dip at the minimum, meaning that $\widehat{m}_{c0}$ is very unforgiving of sub-optimal choices of bandwidth. In this sense, ridged local cubic methods are substantially less robust than skewed local linear ones.

Figure 2.7 reports the pointwise mean squared error (MSE) at the two turning points of m, namely $x_1 = 0.1363$ and $x_2 = \frac{2}{3}$, where $m'(x_j) = 0$ for $j = 1, 2$. These were chosen because they represent points of high curvature, and consequently high bias. They also correspond to sparse and dense design, repsectively. Broadly similar results are observed as in the MISE case. Panels (a) and (b) of Figure 2.7 are the analogues of panel (b) of Figure 2.6, in the cases of x_1 and x_2 respectively. Particularly in the context of sparse design (panel (a) of Figure 2.7), choice of a suboptimal ridge or bandwidth for local cubic methods can have a deleterious effect on pointwise MSE performance. The skewed local linear estimator is more forgiving, with (as in the MISE case) a shallower MSE curve.

Figure 2.8 illustrates the effects, on actual curve estimates, of some of the problems noted above. For each $i = 1, 2, 3$, panels (i.a), (i.b) and (i.c) depict curve estimates, for a typical sample of size $n = 50$, with a different sample for each i, drawn using (a) the "optimally" ridged local cubic estimator $\widehat{m}_{c0}$, (b) the skewed estimator $\widetilde{m}_{\infty}$, and (c) the skewed estimator $\widetilde{m}_{\lambda}$. In each case the curves are stacked in increasing order of the bandwidth used in their construction: $0.50\, h_{opt}$ (for the bottom curve), $0.75\, h_{opt}$, h_{opt} (the long-dashed line), the true curve (unbroken line), $1.25\, h_{opt}$ and $1.5\, h_{opt}$ (the top curve). Here, h_{opt} refers to the bandwidth that is optimal, in a MISE sense, for the respective curve estimator. The problems with numerical instability in local cubic methods, even in some instances for bandwidths larger than h_{opt}, are clear. In particular, the curve estimate is prone to erratic fluctuations in the region of the first mode, indeed at any place where design is relatively sparse. These difficulties are not evident in the case of skewed local linear estimators.

Instability problems with local cubic methods are even more graphic when too-small ridges are employed, as well as suboptimal bandwidths. If the ridges are too large, the estimates fail to properly represent the peaks and troughs in m. For the sake of brevity, figures illustrating these problems are not included.

2.7 Conclusion

In this chapter, we introduce a new bias-reduction technique in nonparametric regression, termed "skewing". We study theoretical properties as well as the finite-sample performance of our estimators. The numerical study shows that when our skewing method is combined with the imputation technique devised by Hall and Turlach (1997b), it proves to be more advantageous than ridged local cubic smoothing. It may be possible to employ the skewing methods to detect discontinuities of the underlying regression curve, but we shall not go into detail here.

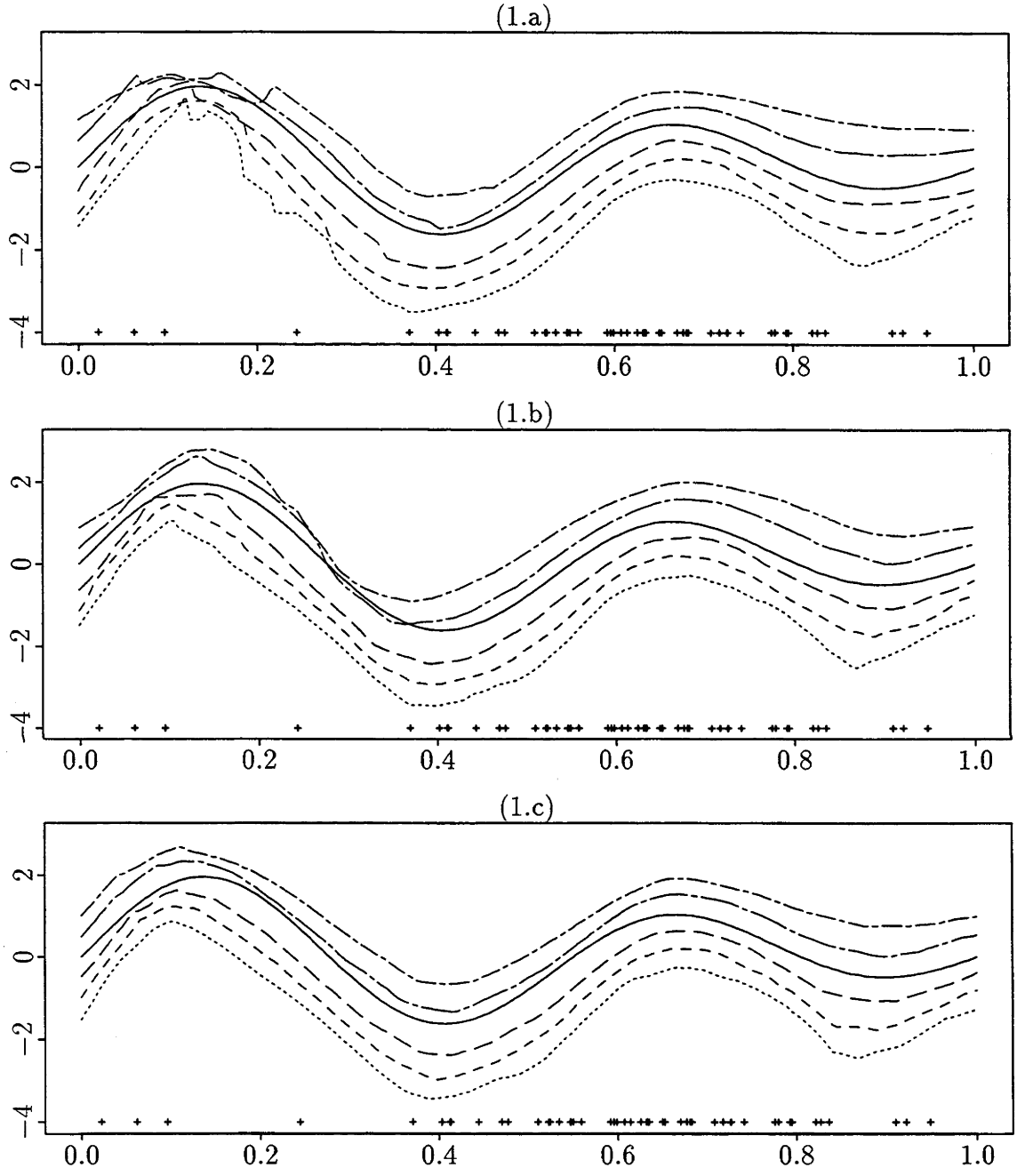


Figure 2.8(i): Typical curve estimates drawn using optimal and suboptimal bandwidths. Three typical data sets were used, represented by the crosses just above the horizontal axis and distinguished by the index $i = 1, 2, 3$ in panels (i.a)–(i.c). Panel (a) depicts ridged local cubic estimators, and the other panels illustrate skewed local linear estimators. See the text for further details.

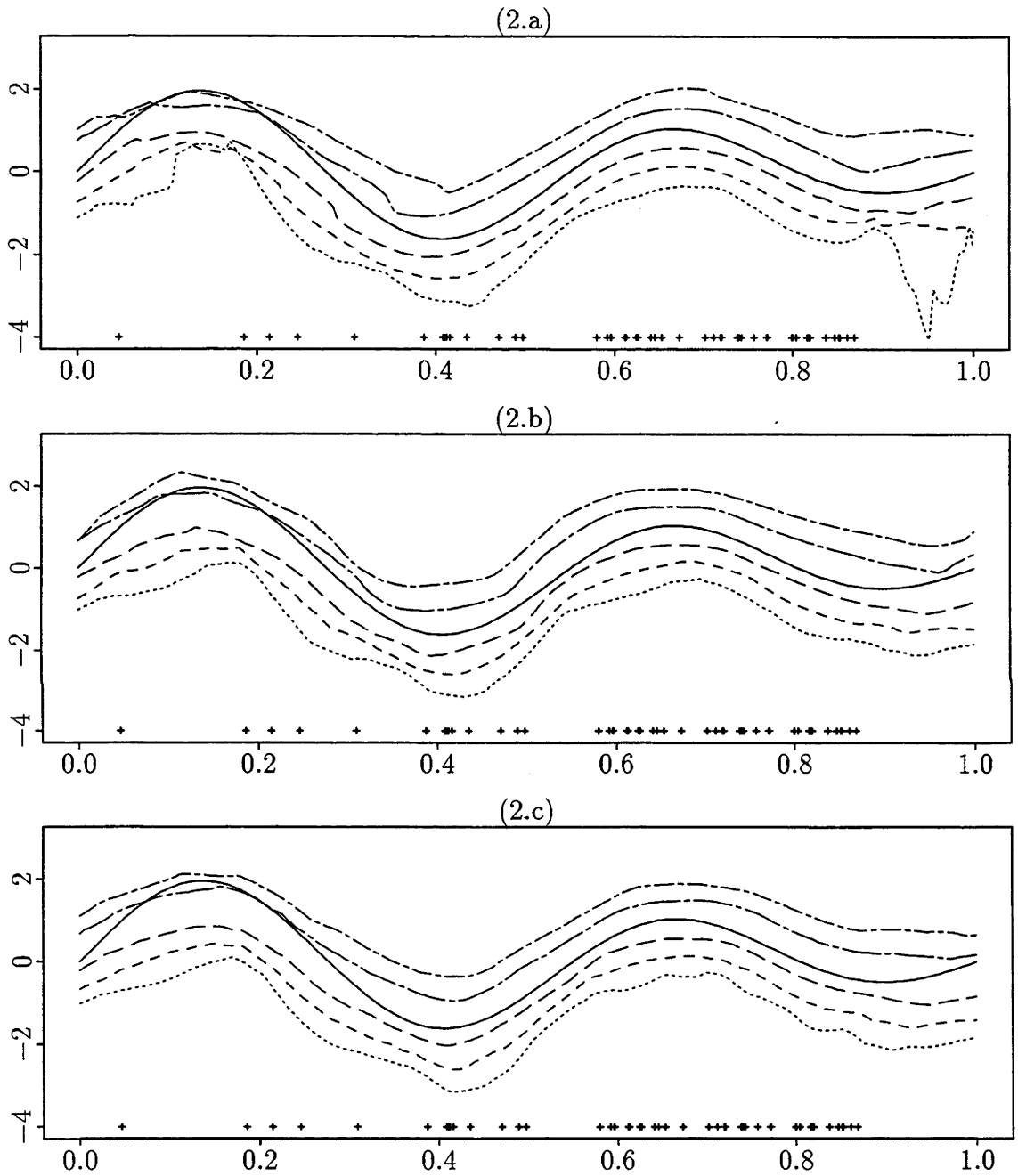


Figure 2.8(ii): Typical curve estimates drawn using optimal and suboptimal bandwidths. See the caption of Figure 2.8(i) for more details.

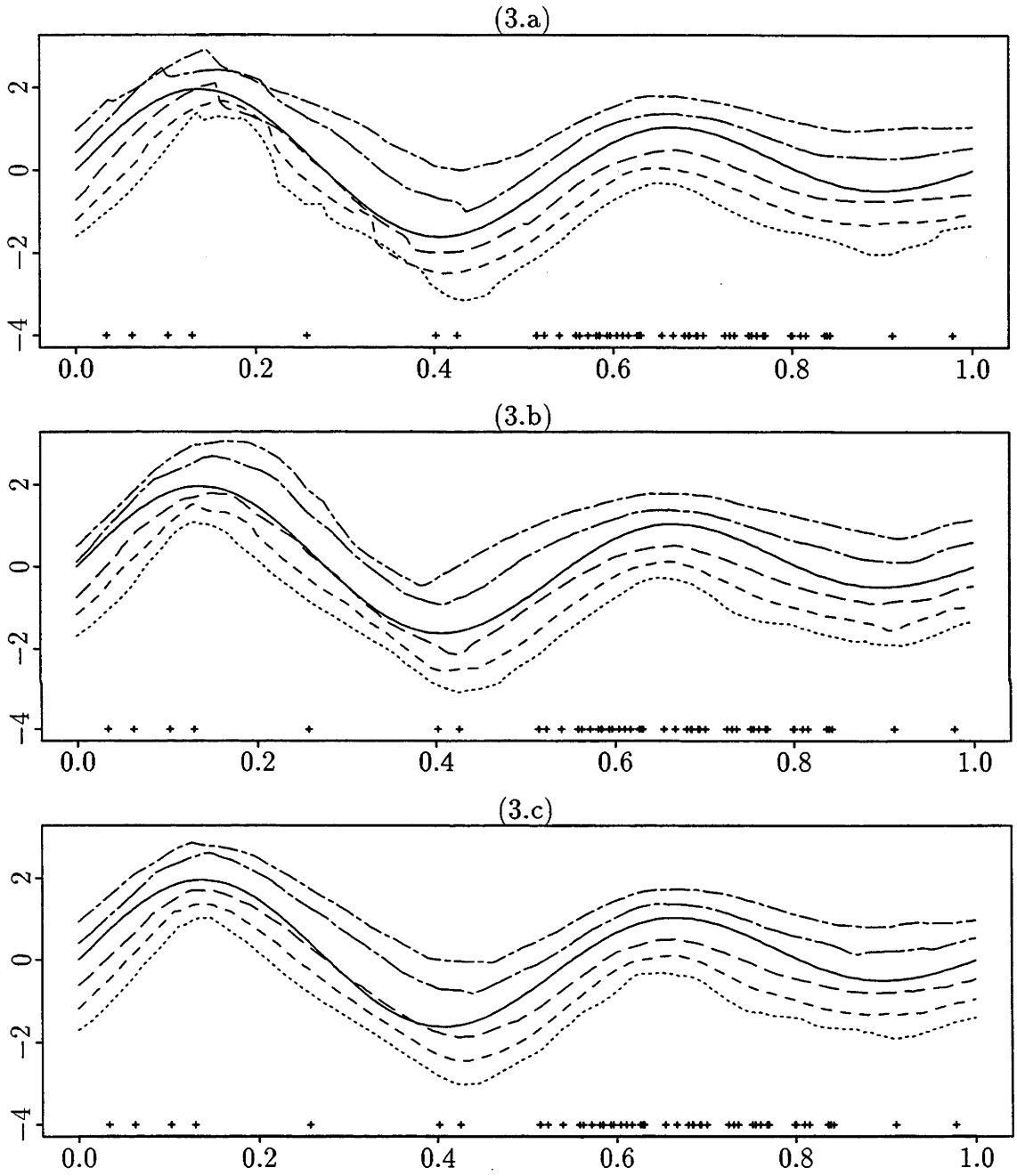


Figure 2.8(iii): Typical curve estimates drawn using optimal and suboptimal bandwidths. See the caption of Figure 2.8(i) for more details.

Chapter 3

Locally Parametric Estimation

3.1 Introduction

So far, we have only concentrated on nonparametric regression problems using local linear smoothing techniques. In this and the following chapters we shall change our focus to density estimation of univariate data using nonparametric methods, which is still in the regime of curve estimation. As in the regression scenario, nonparametric methods for density estimation are very effective in exploring structure in the data. A nonparametric approach does not require a pre-specified parametric model, which is often unavailable in analysing real data. Although a nonparametric estimator may have a slower convergence rate compared with its parametric counterpart, this is only true if we can correctly specify the model. Any wrongly chosen parametric model will lead to an inconsistent density estimator, even if we have plenty of data.

The classical kernel density estimator is one of the most popular and widely-used estimators in practice. Let $X_1, \ldots, X_n$ be a random sample from a density f; let K be a kernel function, taken to be a unimodal density symmetric about 0; and let h be the bandwidth. The classical kernel density estimator is given by

$$\bar{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big). \tag{3.1}$$

The data X_i are weighted around the point x, at which we wish to estimate the density. The kernel K controls the shapes of the weights. Some basic properties of $\bar{f}$ are well-documented in the monographs by Silverman (1986) and Wand and

Jones (1995). The bias and variance of $\bar{f}(x)$ admit the following formulae:

$$E\{\bar{f}(x)\} - f(x) = \tfrac{1}{2}\kappa_2 f''(x)\, h^2 + o(h^2), \tag{3.2}$$

$$\text{var}\{\bar{f}(x)\} = (nh)^{-1}\kappa_1 f(x) + o\{(nh)^{-1}\}, \tag{3.3}$$

where $\kappa_1 = \int K^2$ and $\kappa_2 = \int t^2 K(t)\, dt$. (The notations are consistent with those we used in the previous chapter.)

Locally parametric estimation techniques involve fitting parametric models to the data locally. These techniques have a particularly long history, if one includes among them local linear and local polynomial techniques in nonparametric regression. The recent surge of interest in locally parametric fitting for density and regression estimation is largely motivated by work of Copas (1995), Fan, Farmen and Gijbels (1996), Hjort and Jones (1996) and Loader (1996). The monographs by Wand and Jones (1995) and Fan and Gijbels (1996) should be particularly mentioned.

The density estimators developed by Hjort and Jones (1996), known as locally parametric density estimators, are semiparametric in nature. In their two-parameter form they have theoretical properties comparable to that of $\bar{f}$. In independent work, Loader (1996) also studied local likelihood procedures for density estimation, but restricted attention to log-polynomial models. In this chapter, we shall follow Hjort and Jones' approach by first defining local log-likelihood for density estimation and giving justifications for the approach. Large-sample properties and practical issues will be discussed in later sections. One remarkable result obtained by Hjort and Jones (1996) is that the number of parameters, and not the precise form of the local model, determines the important properties of the estimator. In the two-parameter case, bias and variance of the estimator have the same order as that of the classical kernel estimator when the underlying density f has support on the real line. When estimating f with bounded support, the two-parameter locally-parametric density estimator does not suffer from *boundary bias* problem, which is the discrepancy in the order of bias in the interior and near the boundary, in parallel with the local linear smoother (Fan, 1992).

Other work on semiparametric approaches to density estimation includes that of Olkin and Spiegelman (1987), Copas (1995), Hjort and Glad (1995) and Loader (1996). Olkin and Spiegelman proposed fitting a linear combination of parametric and nonparametric estimates, and defined

$$\hat{f}(x,\omega) = \omega\, g(x,\hat{\theta}) + (1-\omega)\,\bar{f}(x),$$

where $0 \leq \omega \leq 1$ is the weight and is unknown, $g(x, \hat{\theta})$ is a parametric density estimate, and $\bar{f}(x)$ is the usual kernel estimate. This approach unattractively involves the estimation of an extra parameter ω, and complicates to the estimation procedure.

Copas (1995) and Loader (1996) proposed local log-likelihood functions in the context of density estimation, but the methods appear to be less general than Hjort and Jones' (1996) approach. Copas' method has a superficial similarity with Olkin and Spiegelman's (1987) approach, in that his density estimator also spans the continuum between fully parametric and fully nonparametric regimes, but does not require an additional parameter to be estimated. Loader (1996) modelled the logarithm of the density using local polynomials. He showed that his approach may have advantages over kernel methods when estimating tails of densities.

Hjort and Glad (1995) suggested multiplying an initial parametric density estimator $g\{x, \hat{\theta}(x)\}$ by a nonparametric kernel-type estimator of the correction function $r(x) = f(x)/g\{x, \hat{\theta}(x)\}$, say $\hat{r}(x) = (nh)^{-1} \sum_{i=1}^{n} K\{(x - X_i)/h\}/g\{X_i, \hat{\theta}(x)\}$. The estimator $\hat{f}_{HG}(x) = g\{x, \theta(x)\}\hat{r}(x)$, as argued by Hjort and Glad, has the same asymptotic variance and a similar, but often smaller, bias compared with the traditional estimator $\bar{f}(x)$. This method is closely related to the bias-reduction methods proposed by Linton and Nielsen (1994) and Jones, Linton and Nielsen (1995).

In the next chapter we shall show that by incorporating the idea of "skewing", similar to that of local linear smoothing in Chapter 2, to general two-parameter locally-parametric methods for density estimation, one may reduce bias by up to two orders of magnitude and retain variance of the same order.

3.2 Methodology

As in the previous section, let $X_1, \ldots, X_n$ be a random sample from a distribution with density f, which we wish to estimate, and $g(\cdot, \theta)$ be a family of p-parameter functions, indexed by $\theta = (\theta^{(1)}, \ldots, \theta^{(p)})^T$, which we wish to fit to data in a neighbourhood of x. Hjort and Jones (1996) proposed first defining the parameter estimator $\hat{\theta} = \hat{\theta}(x)$ as the maximiser in θ of the local log-likelihood

$$L(x, \theta) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i) \log g(X_i, \theta) - \int K_h(x - t)\, g(t, \theta)\, dt\,, \qquad (3.4)$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$, K is the kernel function (here taken to be a symmetric density), h is the bandwidth. Hjort and Jones (1996), noting similar methods suggested by Copas (1995) and Loader (1996), took their estimator of f to be $\hat{f}(x) = g\{x, \hat{\theta}(x)\}$. Note that $\hat{f}$ need not integrate to one.

In a more general setting, Hjort and Jones considered the parameter estimator $\hat{\theta} = \hat{\theta}(x)$ as the solution in θ of the equation

$$n^{-1} \sum_{i=1}^{n} K_h(x - X_i)\, v_j(x, X_i, \theta) - \int K_h(x - t)\, v_j(x, t, \theta)\, g(t, \theta)\, dt = 0\,, \qquad (3.5)$$

where $v_j(x, t, \theta)$ for $j = 1, \ldots, p$ is a generalised p-parameter weight function. Figure 3.1 gives a graphical example of the locally parametric density estimator constructed using the log-linear model and the standard Normal kernel. The true density has distribution

$$\tfrac{1}{5}N(0,1) + \tfrac{1}{5}N\big(\tfrac{1}{2}, (\tfrac{2}{3})^2\big) + \tfrac{3}{5}N\big(\tfrac{13}{12}, (\tfrac{5}{9})^2\big)\,.$$

One of the very attractive features of the latter approach is its considerable generality, obtained partly through general interpretation of the weight function. For example, if one takes $v_j(x, t, \theta) = (\partial/\partial\theta^{(j)}) \log g(t, \theta)$, the score function of the model, then maximising the likelihood function $L(x, \theta)$ at (3.4) amounts to solving equation (3.5). This is readily seen by differentiating (3.4) with respect to θ and equating to zero. The estimator obtained via this approach is a *local likelihood estimator.*

On the other hand, taking $v_j(x, t, \theta) = (\partial/\partial\theta^{(j)})\, g(t, \theta)$, we obtain a *local least-squares estimator.* To see this, consider the local distance function between $f(\cdot)$ and $g(\cdot, \theta)$ at x, given by

$$d_x(\theta) = \int K_h(x - t)\, \{f(t) - g(t, \theta)\}^2\, dt\,.$$

The $\int K_h(x - t)\, f(t)^2\, dt$ term in $d_x(\theta)$ does not depend on θ, so minimising $d_x(\theta)$ is equivalent to minimising $\int K_h(x - t)\, g(t, \theta)^2\, dt - 2\int K_h(t - x)\, g(t, \theta)\, f(t)\, dt$. The second term depends on the unknown density f, and it may be shown that an unbiased estimator for this term is $2n^{-1}\sum_{i=1}^{n} K_h(x - X_i)\, g(X_i, \theta)$. Hence, we are led naturally to minimising

$$\int K_h(x - t)\, g(t, \theta)^2\, dt - 2n^{-1} \sum_{i=1}^{n} K_h(x - X_i)\, g(X_i, \theta)\,.$$

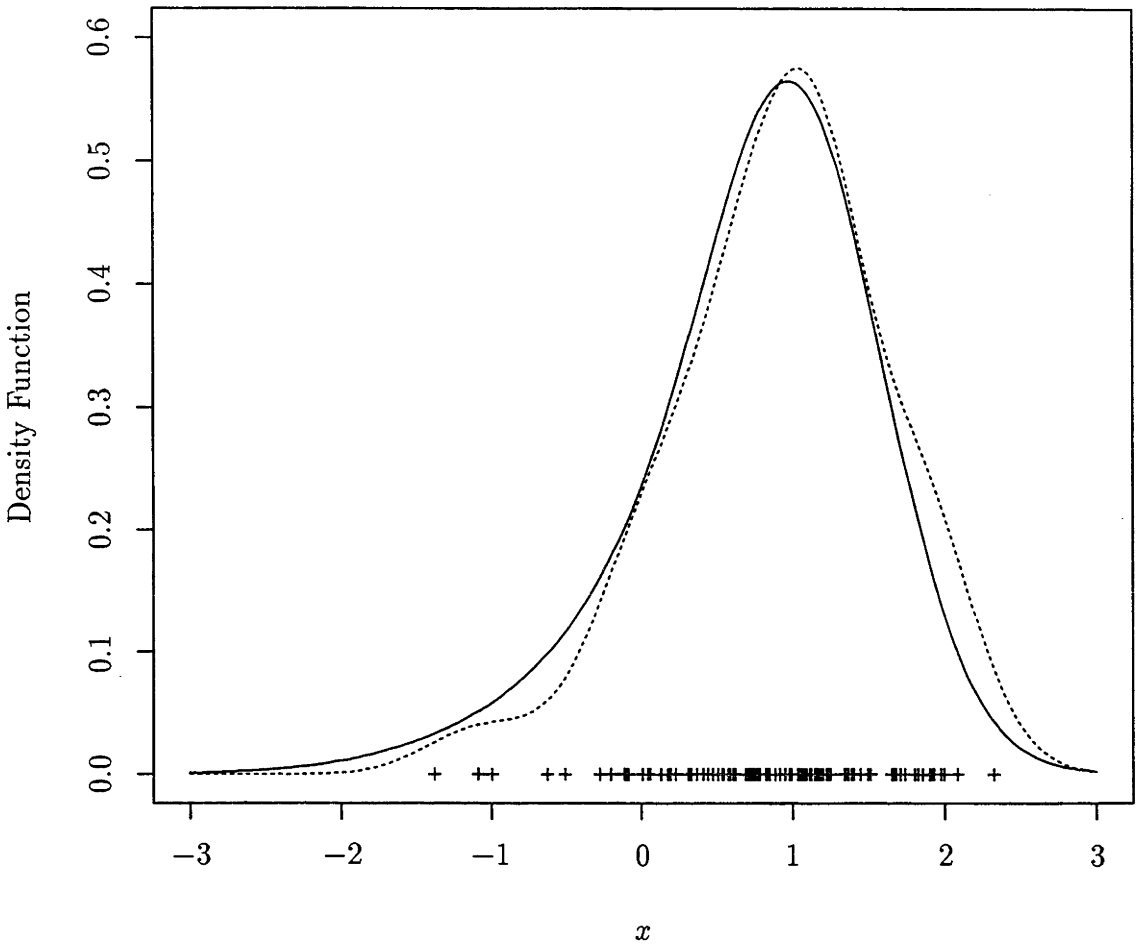


Figure 3.1: An example of the locally parametric density estimator. The sample size is 100, and the crosses above the x-axis represent the data. The solid line depicts the true density, and the dotted line is the locally parametric density estimate. The bandwidth used is 0.310, which minimises the asymptotic MISE.

Differentiating with respect to θ and setting the derivative to zero yields

$$n^{-1}\sum_{i=1}^{n} K_h(x-X_i)\Big(\frac{\partial}{\partial\theta^{(j)}}\Big)\, g(X_i,\theta) \\ -\int K_h(x-t)\left\{\Big(\frac{\partial}{\partial\theta^{(j)}}\Big)\, g(t,\theta)\right\} g(t,\theta)\, dt = 0\,,$$

which is a version of equation (3.5) (obtained by putting $v_j(x,t,\theta) = (\partial/\partial\theta^{(j)})g(t,\theta)$). As the above two examples suggest, it is typically true that $v_j(x,t,\theta)$ does not de-

pend on x.

Note that if h is large in (3.4), $L(x,\theta)$ may be approximated by

$$h^{-1}K(0)\Big\{n^{-1}\sum_{i=1}^{n}\log g(X_i,\theta)-1\Big\},$$

and this reduces to the usual maximum likelihood approach. For small to moderate h the locally parametric estimator utilises primarily local properties of the model, and the method of estimation is essentially nonparametric. In this respect, Hjort and Jones (1996) viewed their locally parametric method as a "continuous bridge" between fully parametric and fully nonparametric options. Simulation results which support this view are given in the numerical section in the next chapter.

As an example of the methodology, consider the local linear model $g(y,\theta) = \theta^{(1)} + (y-x)\theta^{(2)}$. In the context of local likelihood estimation, one seeks a solution of the equation

$$n^{-1}\sum_{i=1}^{n}K_h(x-X_i)\begin{pmatrix}1\\ X_i-x\end{pmatrix} - \int K_h(x-t)\begin{pmatrix}1\\ t-x\end{pmatrix}\{\theta^{(1)}+(t-x)\theta^{(2)}\}\,dt = 0\,.$$

Since $\int K_h(x-t)(t-x)\,dt = 0$, the locally parametric density estimator is

$$\hat f(x) = g\{x,\hat\theta(x)\} = \hat\theta^{(1)} = n^{-1}\sum_{i=1}^{n}K_h(x-X_i)\,,$$

which coincides with the classical kernel estimator.

3.3 Motivations

In this section, we shall follow Hjort and Jones (1996) and give support for the locally parametric method of the previous section, based on a Kullback-Leibler distance argument and a connection to hazard rate estimation in survival data (see also Hjort, 1991, 1997). In the paper by Hjort and Jones, further motivations for their method are given, but we shall not discuss those here.

Definition 3.1. *The Kullback-Leibler distance between two distributions with respective densities f_0 and f_1 is defined as*

$$\ell = \ell(f_0,f_1) = \int f_0(t)\,\log\{f_0(t)/f_1(t)\}\,dt\,.$$

Connections of Kullback-Leibler distance to other distances, e.g. Hellinger or variational distances, are discussed by Reiss (1989, Chapter 3).

Hjort and Jones (1996) defined a local statistical Kullback-Leibler type distance d_K about x between two densities f_0 and f_1 as

$$d_K(f_0, f_1) = \int K_h(x-t)\left[f_0(t) \log \frac{f_0(t)}{f_1(t)} - \{f_0(t) - f_1(t)\} \right] dt\,. \tag{3.6}$$

By expressing $\ell(f_0, f_1)$ as $\int \left[f_0(t) \log\{f_0(t)/f_1(t)\} - \{f_0(t) - f_1(t)\} \right] dt$, and comparing with (3.6), we see that $d_K(f_0, f_1)$ is essentially a locally-weighted version of $\ell(f_0, f_1)$ about x. If we put $f_0 = f$ and $f_1 = g(\cdot, \theta)$ in d_K and minimise over θ, the minimiser $\hat{\theta} = \hat{\theta}(x)$ is the solution of

$$\int K_h(x-t) \left\{ \left(\frac{\partial}{\partial \theta} \right) g(t,\theta) \right\} \left\{ \frac{g(t,\theta) - f(t)}{g(t,\theta)} \right\} dt = 0\,. \tag{3.7}$$

To appreciate the relationship between (3.4) and $d_K\{f, g(\cdot,\theta)\}$, note that on rewriting $L(x,\theta)$ as $\int K_h(x-t) \left\{ \log g(t,\theta)\, dF_n(t) - g(t,\theta)\, dt \right\}$, where F_n is the empirical distribution function based on the sample $X_1, \ldots, X_n$, $L(x,\theta)$ is seen to converge to

$$\int K_h(x-t) \left\{ f(t) \log g(t,\theta) - g(t,\theta) \right\} dt \tag{3.8}$$

in probability as $n \to \infty$. It may be easily shown that $\hat{\theta}(x)$ which solves (3.7) (and hence minimises the local Kullback-Leibler type distance d_K between f and $g(\cdot,\theta)$ about x) also maximises (3.8).

Alternatively, Hjort and Jones (1996) considered the connection with hazard rate estimation in survival analysis. Let $T_1, \ldots, T_n$ denote a sample of survival times from a density $g(t,\theta)$, with distribution $G(t,\theta)$. The complementary function $S(t,\theta) = 1 - G(t,\theta)$ is known as the *survivor function* and $h(t,\theta) = g(t,\theta)/S(t,\theta)$ is the *hazard function*. See for example the monograph by Miller (1981). Let $S_n(t) = 1 - F_n(t)$ be the fraction of the individuals still surviving at time t. Taking $T_0 = 0$, and expressing

$$\begin{aligned} \int S_n(t)\, h(t,\theta)\, dt &= \sum_{i=0}^{n-1} \left(1 - \frac{i}{n}\right) \int_{T_i}^{T_{i+1}} h(t,\theta)\, dt \\ &= -\sum_{i=0}^{n-1} \left(1 - \frac{i}{n}\right) \left\{ \log S(T_{i+1},\theta) - \log S(T_i,\theta) \right\} \end{aligned}$$

$$= -n^{-1} \sum_{i=1}^{n} \log S(T_i, \theta)\,,$$

one obtains the log-likelihood for the hazard model (see also Hjort (1991, 1997)) as

$$\sum_{i=1}^{n} \{ \log h(T_i, \theta) + \log S(T_i, \theta) \} = n^{-1} \int \{ \log h(t, \theta)\, dF_n(t) - S_n(t)\, h(t, \theta)\, dt \}\,.$$

Hjort and Jones argued that the kernel-smoothed local log-likelihood for the model at x, ignoring the factor n^{-1}, is

$$\begin{aligned} &\int K_h(x-t) \{ \log h(t, \theta)\, dF_n(t) - S_n(t)\, h(t, \theta)\, dt \} \\ &\qquad = \int K_h(x-t) \left\{ \log \frac{g(t, \theta)}{S(t, \theta)}\, dF_n(t) - S_n(t)\, \frac{g(t, \theta)}{S(t, \theta)}\, dt \right\}. \end{aligned} \tag{3.9}$$

Replacing $S(t, \theta)$ by the empirical estimate $S_n(t)$ in (3.9) gives

$$\int K_h(x-t) \Big[\{ \log g(t, \theta) - \log S_n(t) \}\, dF_n(t) - g(t, \theta)\, dt \Big]\,. \tag{3.10}$$

Disregarding the term in $\log S_n(t)$, which does not depend on θ, (3.10) has exactly the same form as (3.4) (see also (3.8)). Furthermore, (3.9) converges in probability to

$$\int K_h(x-t) \left\{ f(t) \log \frac{g(t, \theta)}{S(t, \theta)} - S(t)\, \frac{g(t, \theta)}{S(t, \theta)} \right\} dt \tag{3.11}$$

as $n \to \infty$, and Hjort and Jones noted that $\hat{\theta}$ which minimises the local distance

$$\begin{aligned} d\{f, g(\cdot, \theta)\} = \int K_h(x-t) \Bigg[f(t) \left\{ \log \frac{f(t)}{S(t)} - \log \frac{g(t, \theta)}{S(t, \theta)} \right\} \\ - S(t) \left\{ \frac{f(t)}{S(t)} - \frac{g(t, \theta)}{S(t, \theta)} \right\} \Bigg]\, dt\,, \end{aligned}$$

also maximises (3.11) about x. Both connections explicitly motivate the local likelihood function $L(\cdot, \theta)$. Of course, the method may also be motivated from a theoretical viewpoint, which we shall demonstrate in the next section.

3.4 Theoretical Properties

Local linear smoothers enjoy favourable asymptotic properties and boundary behaviour (Fan, 1992, 1993) in nonparametric regression. In density estimation, locally parametric density estimators $\hat{f}$ have similar appealing features. Hjort and

Jones (1996) discussed large-sample properties of locally parametric estimators up to four parameters, but we shall concentrate on two-parameter cases.

In the general two-parameter setting where $g(\cdot, \theta)$ is a family of functions indexed by $\theta = (\theta^{(1)}, \theta^{(2)})^T$, Hjort and Jones (1996) showed that bias and variance of $\hat{f}$ admit the following asymptotic approximations:

$$\begin{aligned} E\{\hat{f}(x)\} - f(x) &= g\{x, \theta_0(x)\} - f(x) + O\{(nh)^{-1}\} \\ &= \tfrac{1}{2}\kappa_2 h^2 \big[f''(x) - g''\{x, \theta_0(x)\}\big] + O\big\{h^4 + (nh)^{-1}\big\}, \quad (3.12) \\ \operatorname{var}\{\hat{f}(x)\} &= (nh)^{-1}\kappa_1 f(x) + o\big\{(nh)^{-1}\big\}, \quad (3.13) \end{aligned}$$

as $h \to 0$ and $nh \to \infty$, where $\kappa_1 = \int K^2$, $\kappa_2 = \int t^2 K(t)\, dt$, $g^{(j)}(x, \theta)$ (or g with j dashes) denotes $(\partial/\partial x)^j g(x, \theta)$, and $\theta_0(y) = \theta_0(y, h)$ is the solution in θ of the equation

$$\int K_h(y - t)\, v_j(y, t, \theta)\, \{f(t) - g(t, \theta)\}\, dt = 0 \quad \text{for} \quad j = 1, 2. \qquad (3.14)$$

Note that this equation is a "population version" of (3.5). We assume that, for each y and all sufficiently small h, $\theta_0(y)$ exists and is unique. Other regularity conditions for (3.12) and (3.13) will be discussed in the next chapter.

To see the development of (3.12), Hjort and Jones pointed out that since the left-hand side of (3.5) has mean zero on substituting $\theta = \theta_0(x)$, then under mild regularity conditions (Shao, 1991), we have

$$\hat{\theta}(x) = \theta_0(x) + O_p\big\{(nh)^{-1}\big\},$$

for fixed h and large n; and by the delta method,

$$\hat{f}(x) = g\{x, \theta_0(x)\} + O_p\big\{(nh)^{-1}\big\}. \qquad (3.15)$$

As we decrease h, using (3.14) and Taylor-expansion, we may show that

$$\begin{aligned} f(x) - g\{x, \theta_0(x)\} &= \tfrac{1}{2}\kappa_2 h^2 \Big(g''\{x, \theta_0(x)\} - f''(x) \\ &\qquad + 2\, \frac{v_j'\{x, x, \theta_0(x)\}}{v_j\{x, x, \theta_0(x)\}} \big[g'\{x, \theta_0(x)\} - f'(x) \big] \Big) \\ &\qquad + O(h^4) \quad \text{for} \quad j = 1, 2, \qquad (3.16) \end{aligned}$$

where v_j' denotes $(\partial/\partial t) v_j(x, t, \theta)$. Assuming that v_1 and v_2 are functionally independent and satisfy certain regularity assumptions, (3.16) implies $g'\{x, \theta_0(x)\} - f'(x) =$

$o(1)$ as $h \to 0$. (In fact, from (4.17) in the proof of Theorem 4.1, $g'\{x, \theta_0(x)\} - f'(x) = O(h^2)$.) Combining this result, (3.15) and (3.16) gives the bias expression (3.12).

The derivation of (3.13) is comparatively difficult. However, if $g(\cdot, \theta)$ and v_j can be suitably reparametrized, much simplification can be gained as demonstrated in Section 4.2 of Hjort and Jones (1996). They showed that, if we may assume the reparametrized $v_j(x, t, \theta)$ functions are of the form $c_1 + c_2(ht) + c_3(ht)^2 + \ldots$ for small h, then

$$\operatorname{var}\{\hat{f}(x)\} = (nh)^{-1}\,\tau(K)^2\,f(x) + o\{(nh)^{-1}\},$$

where $\tau(K)^2 = \omega^T M_1^{-1} M_2 M_1^{-1} \omega$, with $\omega^T = (1, 0), M_1 = \operatorname{diag}(1, h^2\kappa_2), M_2 = \operatorname{diag}(\kappa_1, h^2\kappa_3)$ and $\kappa_3 = \int t^2 K(t)^2\,dt$; whence it follows that $\tau(K)^2 \sim \kappa_1$.

The analysis we have discussed so far assumes the underlying density f has support on the whole real line. Suppose now that f has a left-hand boundary, the location of which is known at $x = 0$. We are interested in estimating f at $x = \alpha h$, where $0 \le \alpha < 1$. For the classical kernel density estimator $\bar{f}$, boundary bias takes the form:

$$\begin{aligned} E\{\bar{f}(x)\} - f(x) &= \int_{-1}^{\alpha} K(t) f(x - ht)\,dt - f(x) \\ &= \{\nu_0(\alpha) - 1\}\,f(x) - h\,\nu_1(\alpha)\,f'(x) + O(h^2), \end{aligned}$$

where $\nu_k(\alpha) = \int_{-1}^{\alpha} t^k\,K(t)\,dt$ and K has support confined to $[-1, 1]$. Since $\nu_0(\alpha)$ does not typically equal one, $\bar{f}$ is not even a consistent estimator, and the boundary density is usually underestimated. To correct for this problem one may consider normalising $\bar{f}$ by dividing by $\nu_0(\alpha)$ to achieve $O(h)$ bias and consistency. To further reduce bias, Gasser and Müller (1979) suggested the use of modified kernels, or so-called "boundary kernels", to alleviate the impact of boundary effects and to attain $O(h^2)$ bias everywhere. One simple class of boundary kernels is

$$\widetilde{K}_\alpha(t) = \left\{\frac{\nu_2(\alpha) - \nu_1(\alpha)\,t}{\nu_0(\alpha)\nu_2(\alpha) - \nu_1(\alpha)^2}\right\} K(t)\,I(-1 \le t \le \alpha). \tag{3.17}$$

It is easy to check that $\widetilde{K}_\alpha$ satisfies $\int \widetilde{K}_\alpha = 1$ and $\int t\,\widetilde{K}_\alpha(t)\,dt = 0$. An account of boundary kernels can be found in Müller (1991).

In terms of boundary bias, $\hat{f}$ is clearly superior to the traditional estimator $\bar{f}$. Hjort and Jones (1996) showed that bias and variance of the two-parameter locally-parametric estimator $\hat{f}$ admit the following asymptotic expansions:

$$E\{\hat{f}(x)\} - f(x) \sim \tfrac{1}{2} Q(\alpha)\,[f''(x) - g''\{x, \theta_0(x)\}]\,h^2, \tag{3.18}$$

$$\operatorname{var}\{\hat{f}(x)\} \sim (nh)^{-1}\,R(\alpha)\,f(x), \tag{3.19}$$

where

$$Q(\alpha) = \frac{\nu_2(\alpha)^2 - \nu_1(\alpha)\,\nu_3(\alpha)}{\nu_0(\alpha)\,\nu_2(\alpha) - \nu_1(\alpha)^2}, \qquad R(\alpha) = \frac{\int \{\nu_2(\alpha) - \nu_1(\alpha)\,t\}^2\,K(t)^2\,dt}{\left\{\nu_0(\alpha)\,\nu_2(\alpha) - \nu_1(\alpha)^2\right\}^2}.$$

Putting $\alpha = 1$ in (3.18) and (3.19), the asymptotic bias and variance terms are equivalent to those at (3.12) and (3.13) respectively. Thus, similarly to local linear smoothing in nonparametric regression, the two-parameter, locally-parametric density estimator does not suffer from boundary bias problem (Fan, 1992). Note that this result depends solely on the number of parameters fitted, not at all on the choice of model. The variance terms at (3.13) and (3.19) have no dependence on the local model $g(\cdot, \theta)$ and the weight functions v_j. The bias terms, on the other hand, depend on the model through the factor $f'' - g''(\cdot, \theta)$, which corresponds to the difference between the second derivative of f and the local curvature of the model. Further bias reduction may be possible if we choose $g(\cdot, \theta)$ appropriately to minimise this second order difference. Hjort and Jones (1996) conjectured that for $k \geq 1$: (i) $(2k-1)$- or $(2k)$-parameter locally parametric estimation affords $O(h^{2k})$ bias; (ii) boundary bias has order $O(h^k)$ for k-parameter fitting.

As an illustration in the context of density estimation at a boundary, consider the local linear model $g(y, \theta) = \theta^{(1)} + (y - x)\theta^{(2)}$ in Section 3.2. To work out the locally parametric density estimator at $x = \alpha h$, we have to solve for $\theta^{(1)}$ in the following equations:

$$\begin{aligned} n^{-1}\sum_{i=1}^{n} K_h(x - X_i) - \nu_0(\alpha)\theta^{(1)} - \nu_1(\alpha)\theta^{(2)} &= 0, \\ n^{-1}\sum_{i=1}^{n} K_h(x - X_i)(X_i - x) - \nu_1(\alpha)\theta^{(1)} - \nu_2(\alpha)\theta^{(2)} &= 0. \end{aligned}$$

Simple algebra gives

$$\hat{f}(x) = g\{x, \hat{\theta}(x)\} = n^{-1}\sum_{i=1}^{n} \frac{K_h(x - X_i)\,\nu_2(\alpha) - K_h(x - X_i)(X_i - x)\,\nu_1(\alpha)}{\nu_0(\alpha)\,\nu_2(\alpha) - \nu_1(\alpha)^2}.$$

Note that this is equivalent to the kernel estimator employing the modified boundary kernel at (3.17).

3.5 Practical Issues

So far we have only focussed on theoretical aspects of the locally parametric method. To put the method into practice in the context of local likelihood estimation, one has to choose a parametric model $g(\cdot, \theta)$ and solve the equation at (3.5) for θ (putting $v_j(x, t, \theta) = (\partial/\partial\theta^{(j)}) \log g(t, \theta)$). For some choice of kernel functions and local models, explicit solution of (3.5) may be possible, and the density estimator may be given in closed-form. A simple example has already been given in Section 3.2. See also Section 4.2 when a log-linear model and standard Normal kernel are employed.

When no explicit solution is available, which is often the case, one has to resort to numerical methods by either maximising the local log-likelihood $L(x, \theta)$ for each x at (3.4), subject to constraints on the parameters of the model; or solving (3.5) for each x (existence and uniqueness of solution is ensured if the local log-likelihood is concave) by iterative methods. Computational issues arise both for the efficiency of the method and the accuracy of the estimates. Although increasing the number of parameters in the model leads to further bias reduction, the practical gain from using high-order methods is not clear-cut, as pointed out by Hjort and Jones (1996). This is analogous to using a high-order kernel to achieve bias reduction where the asymptotic advantages may not readily pass on in finite samples (Marron and Wand, 1992). Moreover, increasing the number of parameters poses greater computational challenges and may create numerically instability, owing to the complexity of the method. Therefore, two-parameter models seem to be a natural choice, not least because of their favourable theoretical properties compared with one-parameter models, and their computational simplicity compared with high-order cases.

Chapter 4

Skewing in Density Estimation

4.1 Introduction

In the last chapter, we reviewed aspects of the locally parametric density estimator following the Hjort and Jones' (1996) approach. By employing a "skewing" technique similar to that introduced in the context of local linear regression in Chapter 2, we shall show that the bias of general locally parametric methods can be reduced by up to two orders of magnitude while retaining the variance to the same order of magnitude.

Skewing methods for regression discussed in Chapter 2 involved first calculating the usual local linear smoother at a point x' that is a short distance from the place x where we wish to estimate the regression function, and then evaluating this approximation at x. We showed that $x' - x$ depends in a very simple way on the bandwidth and the kernel. Using skewing in this simple form reduces bias by one order of magnitude, and incurs only a moderate increase in variance. Taking the average of two such estimators computed at either side of x reduces bias by another order of magnitude, still at the expense of only a constant-factor inflation of variance. The extension of these skewing techniques to locally parametric density estimation methods is the main theme of this chapter.

There exists a variety of bias-reduction techniques for remedying inadequacies of the classic kernel density estimator, and higher-order kernel methods are arguably the most commonly used in practice (Marron and Wand, 1992; Jones and Foster, 1993). Nevertheless, higher-order kernels take on negative values, and may result in negative density estimates. This increases difficulties in terms of interpretability and

plausibility of the method. Skewed estimators are, on the other hand, guaranteed to be nonnegative, since they are convex combinations of evaluations of a nonnegative parametric function $g(\cdot, \theta)$. Therefore, skewing reproduces the bias-reduction effect of higher-order density estimation, without risking the occurrence of negative estimates. A brief and recent account of bias-reduction methods in density estimation may be found in Jones, Linton and Nielsen (1995) and Jones and Signorini (1997).

The improvements in performance are available for general kernels and general approaches to locally parametric estimation, for example those based on either local likelihood or local least squares discussed in Section 3.2. They allow mean squared error to be reduced from order $n^{-4/5}$, in the case of standard kernel or locally parametric methods, up to order $n^{-8/9}$ for certain forms of skewed estimators.

We have followed the development of Hjort and Jones (1996) in our presentation and discussion, which applies in a particularly broad setting. High-order methods in curve estimation include work of Ruppert and Wand (1994), in the context of local high-order polynomial modelling in regression, as well as contributions by Hjort and Jones (1996) and Loader (1996) to high-order local log-polynomial modelling in density estimation.

This chapter is organised as follows. Sections 4.2 and 4.3 introduce our skewing methods, and Section 4.4 gives possible extensions of the methods to general curve estimation. Theoretical aspects of our skewed estimators are discussed in Sections 4.5 and 4.6, and finite-sample behaviour of the estimators is addressed through a simulation study in Section 4.7.

4.2 Skewing

Let $X_1, \ldots, X_n$ be independent and identically distributed with density f. Recall that in the general two-parameter locally-parametric density estimation described in Section 3.2, one solves the equation

$$n^{-1} \sum_{i=1}^{n} K_h(x - X_i)\, v_j(x, X_i, \theta) - \int K_h(x - t)\, v_j(x, t, \theta)\, g(t, \theta)\, dt = 0 \qquad (4.1)$$

in θ around each x where we wish to estimate the underlying density. Here, $g(\cdot, \theta)$ is a family of two-parameter functions, indexed by $\theta = (\theta^{(1)}, \theta^{(2)})^T$; $K_h(\cdot) = h^{-1}K(\cdot/h)$, K is the kernel function, taken to be either the Standard Normal density or a symmetric, compactly supported, unimodal density; h is the bandwidth; and $v_j(x, t, \theta)$,

for $j = 1, 2$, is a generalised two-parameter weight function. The locally parametric density estimator of f is taken to be $\hat{f}(x) = g\{x, \hat{\theta}(x)\}$, where $\hat{\theta} = \hat{\theta}(x)$ is the solution of (4.1).

Following standard practice in local curve fitting, Hjort and Jones (1996) (and others working on locally parametric methods) computed $\hat{f}$ symmetrically. In other words, they weighted the data on either side of x symmetrically and calculated $\hat{f}$ at the "centre" of the weights. Skewing involves using symmetric weights at an off-centre point x', but nevertheless calculating the estimator at x. Thus, we replace $\hat{f}(x) = g\{x, \hat{\theta}(x)\}$ by $\hat{f}(x|x') = g\{x, \hat{\theta}(x')\}$.

As an example of skewing, let $g(\cdot, \theta)$ be the log-linear model such that $g(y, \theta) = \theta^{(1)} \exp\{(y - x)\theta^{(2)}\}$, and

$$v_j(x, t, \theta) = \Big(\frac{\partial}{\partial \theta^{(j)}}\Big) \log g(t, \theta) = \Big(\frac{\partial}{\partial \theta^{(j)}}\Big) \left\{ \log \theta^{(1)} + (t - x)\, \theta^{(2)} \right\}.$$

Then $\hat{\theta}^{(1)}, \hat{\theta}^{(2)}$ are solutions of the equations

$$\begin{aligned} n^{-1} \sum_{i=1}^{n} K_h(x - X_i)(\theta^{(1)})^{-1} - \int K_h(x - t) \exp\left\{(t - x)\, \theta^{(2)}\right\} dt &= 0, \\ n^{-1} \sum_{i=1}^{n} K_h(x - X_i)\,(X_i - x) - \int K_h(x - t)\,(t - x)\, \theta^{(1)} \exp\left\{(t - x)\theta^{(2)}\right\} dt &= 0. \end{aligned}$$

Define ψ to be the moment generating function corresponding to the density K, let $\psi(s) = \int \exp(ts)\, K(t)\, dt$, and put $\bar{f}_k(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i)\,(X_i - x)^k$. Note that $\bar{f} = \bar{f}_0$, the classical kernel density estimator. In this notation, the above system of equations may be rewritten as $\bar{f}_0 = \hat{\theta}^{(1)} \psi(h\hat{\theta}^{(2)})$ and $\bar{f}_1 = h\hat{\theta}^{(2)} \psi'(h\hat{\theta}^{(2)})$; and

$$\hat{f}(x|x') = g\big\{x, \hat{\theta}(x')\big\} = \bar{f}_0(x)\, \psi\big\{h\hat{\theta}^{(2)}(x')\big\}^{-1} \exp\left\{(x - x')\, \hat{\theta}^{(2)}(x')\right\}. \tag{4.2}$$

When K is the Standard Normal kernel, we have $\psi(t) = \exp(t^2/2)$, and

$$\bar{f}'(x) = (nh^3)^{-1} \sum_{i=1}^{n} (X_i - x) \exp\{-(x - X_i)^2/(2h^2)\} = h^{-2} \bar{f}_1(x).$$

So, $\bar{f}_1/\bar{f}_0 = h^2 \hat{\theta}^{(2)}$, and for each constant c,

$$\begin{aligned} \hat{f}(x|x + ch) &= \bar{f}(x + ch)\, \psi\big[\bar{f}_1(x + ch)\{h f_0(x + ch)\}^{-1}\big]^{-1} \\ &\qquad \times \exp\big[- c\bar{f}_1(x + ch)\{h f_0(x + ch)\}^{-1}\big] \\ &= \bar{f}(x + ch) \exp\left[\tfrac{1}{2}c^2 - \tfrac{1}{2}\big\{c + h\bar{f}'(x + ch)\, \bar{f}(x + ch)^{-1}\big\}^2\right]. \end{aligned} \tag{4.3}$$

Taking $c = 0$ in (4.3) gives the local log-linear density estimator of Hjort and Jones (1996) and Loader (1996).

Hjort and Jones commented that in this example, when $c = 0$ the parameter estimate $\hat{\theta}^{(2)}$ is "only somewhat silently present". That cannot be said of the case $c \neq 0$ in which we are interested. Those authors also argued that $\hat{\theta}^{(2)}$ might be computed separately from $\hat{\theta}^{(1)}$, using a larger bandwidth or post-smoothing the values of $\hat{\theta}^{(2)}$ before plugging into (4.2). Following those prescriptions here would destroy the bias-reduction properties of estimators constructed by skewing.

4.3 Skewed Estimators and Their Properties

Using the general setting described in the previous section, we shall give here a list of skewed estimators which have favourable bias properties. Recall that in skewing, we put symmetric weights around a point x' that is slightly to one side of x, at which point we wish to compute the estimator $\hat{f}(x|x') = g\{x, \hat{\theta}(x')\}$. Indeed, choosing $x' = x_\pm \equiv x \pm \kappa_2^{1/2} h$ (for either choice of the + and − signs), where $\kappa_2 = \int t^2 K(t)\, dt$, produces estimators $\tilde{f}_\pm(x) = g\{x, \hat{\theta}(x_\pm)\}$ whose biases are $O\{h^3 + (nh)^{-1}\}$ rather than $O\{h^2 + (nh)^{-1}\}$. This result, together with the ones mentioned below, will be derived in Section 4.6. Using the symmetric convex combination $\tilde{f} = \frac{1}{2}(\tilde{f}_+ + \tilde{f}_-)$ reduces bias by another order of magnitude, to $O\{h^4 + (nh)^{-1}\}$. More generally, employing the estimator

$$\tilde{f}_\lambda(x) = (2\lambda + 1)^{-1} \left\{ \lambda \hat{f}(x|x + lh) + \hat{f}(x|x) + \lambda \hat{f}(x|x - lh) \right\},$$

where $0 \leq \lambda < \infty$ and

$$l = l(\lambda) = \left\{ (1 + 2\lambda)\, \kappa_2 / (2\lambda) \right\}^{1/2}, \tag{4.4}$$

also reduces bias to $O\{h^4 + (nh)^{-1}\}$. (Note that $\tilde{f} = \tilde{f}_\infty$.) Thus, we have

$$E(\tilde{f}_\pm) = f + O\left\{h^3 + (nh)^{-1}\right\}, \quad E(\tilde{f}) = f + O\left\{h^4 + (nh)^{-1}\right\},$$

$$E(\tilde{f}_\lambda) = f + O\left\{h^4 + (nh)^{-1}\right\}. \tag{4.5}$$

The variance remains at order $(nh)^{-1}$ throughout these manipulations. Indeed, under regularity conditions implicit in Hjort and Jones (1996) (see for example (4.9) below),

$$\mathrm{var}\,(\tilde{f}_\pm) \sim (nh)^{-1} \left(\kappa_1 + \kappa_2^{-1} \kappa_3\right) f, \quad \mathrm{var}\,(\tilde{f}_\lambda) \sim (nh)^{-1}\, V(\lambda)\, f, \tag{4.6}$$

as $h \to 0$ and $n \to \infty$ in such a manner that $nh \to \infty$, where $\kappa_1 = \int K^2$, $\kappa_3 = \int t^2 K(t)^2\,dt$ and

$$\begin{aligned} V(\lambda) &= (2\lambda+1)^{-2}\Big[\big(2\lambda^2+1\big)\kappa_1 + \big(6\lambda+1\big)\int K(u-l)\,K(u)\,du \\ &\quad +\tfrac{1}{2}\big(4\lambda+1\big)^2\int K(u-l)\,K(u+l)\,du \\ &\quad +\lambda(2\lambda+1)\,\kappa_2^{-1}\int u^2\big\{K(u)^2 - K(u-l)\,K(u+l)\big\}\,du\Big]. \quad (4.7) \end{aligned}$$

Formula (4.6) for $\mathrm{var}\,(\tilde{f}_\lambda)$ holds when $\lambda = \infty$, so that $\mathrm{var}\,(\tilde{f}) \sim (nh)^{-1}V(\infty)\,f$.

These are the same variances that arise in skewed linear approximation in nonparametric regression in Chapter 2. That is to be expected, given the interpretation of nonparametric density estimation as regression with Poisson-distributed errors (see also Section 2.3 of Hjort and Jones (1996)). The size of $V(\lambda)$, for $0 \le \lambda \le \infty$, has already been discussed in Chapter 2. Depending on the choice of K and λ, $V(\lambda)$ can actually be smaller that κ_1, although it is generally a little larger.

Continuing the example in Section 4.2, we see that the estimators $\tilde{f}_+$ and $\tilde{f}_-$ of f given by

$$\tilde{f}_\pm(x) = \bar{f}(x \pm h)\exp\Big[\tfrac{1}{2} - \tfrac{1}{2}\big\{1 \pm h\bar{f}'(x \pm h)\bar{f}(x \pm h)^{-1}\big\}^2\Big], \qquad (4.8)$$

where the $+$ and $-$ signs are chosen respectively, have bias of size h^3. These are obtained by putting $c = 1$ in (4.3). Moreover, the estimator $\tilde{f} = \frac{1}{2}\,(\tilde{f}_+ + \tilde{f}_-)$ has bias of size h^4, and each of $\tilde{f}_+, \tilde{f}_-, \tilde{f}$ has variance of size $(nh)^{-1}$. By way of comparison, $\bar{f}$ itself has variance of size $(nh)^{-1}$ but larger bias, of size h^2.

While the estimator at (4.3) was derived in the special case of the Standard Normal kernel, it is appropriate much more generally. Indeed, taking $\bar{f}$ to be a general kernel estimator computed using a kernel with $\kappa_2 = 1$ (where, here and in the remainder of this paragraph, κ_j is interpreted for the kernel used to compute $\bar{f}$), and putting $c = \pm 1$, the estimator $\check{f}_\pm(x) = \hat{f}(x|x \pm ch)$ (with the right-hand side given by (4.3)) satisfies $E(\check{f}_\pm) = f + O\{h^3 + (nh)^{-1}\}$ and $\mathrm{var}\,(\check{f}_\pm) \sim (nh)^{-1}(\kappa_1 + \kappa_3)\,f$. This is the analogue of (4.5) and (4.6) (taken there for $\tilde{f}_\pm$) in the case of $\check{f}_\pm$. These results may be derived after little more than Taylor expansion. To see the development of bias expansion, first note that

$$E\big[\bar{f}'(x+ch)\{\bar{f}(x+ch)\}^{-1}\big] = f(x)^{-1}\big[f'(x)+ch\{f''(x)-f(x)^{-1}f'(x)^2\}\big]+O(h^2)\,,$$

using the fact that $E(\bar{f}') = f' + \frac{1}{2}\kappa_2 h^2 f''' + O(h^3)$, and hence

$$\begin{aligned} E\{\check{f}_\pm(x)\} &= \{f(x\pm ch)+\tfrac{1}{2}h^2\kappa_2 f''(x\pm ch)+O(h^3)\} \\ &\quad\times\exp\Big(\mp chf(x)^{-1}\big[f'(x)\pm ch\{f''(x)-f(x)^{-1}f'(x)^2\}\big] \\ &\quad\quad -\tfrac{1}{2}h^2 f(x)^{-2}f'(x)^2+O(h^3)\Big) \\ &= \{f(x)\pm chf'(x)+\tfrac{1}{2}h^2f''(x)(c^2+\kappa_2)+O(h^3)\} \\ &\quad\times\Big[1\mp chf(x)^{-1}f'(x)-(ch)^2f(x)^{-1}\{f''(x)-f(x)^{-1}f'(x)^2\} \\ &\quad\quad +\tfrac{1}{2}h^2f(x)^{-2}f'(x)^2(c^2-1)\Big]+O(h^3)\,, \end{aligned}$$

choosing the + and − signs respectively. It is readily seen that the term involving h vanishes, and that the coefficient of the h^2 term is $\frac{1}{2}h^2\{(\kappa_2-c^2)f''(x)+(c^2-1)f(x)^{-1}f'(x)^2\}$, which equals zero on choosing $c = \pm\kappa_2^{1/2} = \pm 1$. Likewise, one may derive versions of (4.5) and (4.6), for linear combinations of estimators such as $\check{f}_\pm$. This gives rise to analogues $\check{f}$ and $\check{f}_\lambda$ of $\tilde{f}$ and $\tilde{f}_\lambda$. Similarly, versions of (4.3) that arise for kernels other than the Normal may be shown to produce a variety of new estimators which enjoy good bias-reduction properties, provided the kernel is sufficiently smooth to allow the necessary Taylor expansion needed in the argument.

There are other versions of skewed estimators which can reduce bias by orders of magnitude and retain variance of the same order (see also Section 2.5). An example is

$$\lambda\,\hat{f}(x|x)+(1-\lambda)\,\hat{f}(x|x+lh)\,,$$

where $0<\lambda<1$. The choice of $l = \pm\{\kappa_2/(1-\lambda)\}^{1/2}$ reduces bias by an order of magnitude. This result can be readily obtained from the proof of Theorem 4.1.

4.4 Extensions to General Curve Estimation

Apart from density estimation, locally parametric methods have been employed in a variety of curve estimation problems. In the context of generalised linear models, local likelihood methods are used to estimate local parameters. See Sections 5.3 and

5.4 of the monograph by Fan and Gijbels (1996) for details. The skewing idea is also applicable to these settings, and may be shown to reduce asymptotic bias.

We shall illustrate the main ideas using quasi-likelihood models (see McCullagh and Nelder, 1989, Chapter 9). Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample with conditional mean $\mu(x) = E(Y|X = x)$, which we wish to estimate. Quite often, we may not have sufficient knowledge to specify a likelihood function, but we may be able to spell out relationships between the mean and variance. We assume that $V\{\mu(x)\} = \mathrm{var}\,(Y|X = x)$ for some known function V. The function $q(\mu, Y) = \{V(\mu)\}^{-1}(Y - \mu)$ enjoys properties similar to a log-likelihood derivative:

$$\begin{aligned} E\{q(\mu, Y)\} &= 0\,, \\ \mathrm{var}\,\{q(\mu, Y)\} &= \{V(\mu)\}^{-1}\,, \\ -E\left\{\frac{\partial q(\mu, Y)}{\partial \mu}\right\} &= \{V(\mu)\}^{-1}\,. \end{aligned}$$

It follows that the *quasi-likelihood* function, defined as

$$Q(\mu, y) = \int_{\mu} q(t, y)\, dt\,,$$

behaves analogously to the usual log-likelihood function.

In parametric generalised linear models we normally model a linear transformation of the conditional mean, say $\eta(x) = g\{\mu(x)\}$ where $\eta(x) = \alpha + \beta x$. The function g is known as a *link* function and is pre-specified. Instead of taking η as a linear function throughout the entire region of interest, we may assume that η has sufficient local smoothness to be approximated locally by a linear function. In other words, we assume $\eta(u) \approx \alpha + \beta(u - x)$ in the neighbourhood of x at which we wish to estimate the conditional mean. Estimating $\eta(x)$ by local quasi-likelihood methods involves finding α and β to maximise

$$\sum_{i=1}^{n} Q\left[g^{-1}\{\alpha + \beta(X_i - x)\}, Y_i\right] K_h(X_i - x)\,.$$

Let $\hat{\alpha}(x)$ and $\hat{\beta}(x)$ be the maximisers, and put $\hat{\eta}(u|x) = \hat{\alpha}(x) + \hat{\beta}(x)(u - x)$. Then, $\hat{\eta}(x|x)$ corresponds to the usual classic local linear estimator, with asymptotic bias of size h^2 and asymptotic variance of size $(nh)^{-1}$ (Fan, Heckman and Wand, 1995). Note that, although the selection of a link function g becomes less crucial in the nonparametric setting, it is still advisable to choose a link which ensures that the

estimator is range-preserving and that the quasi-likelihood is convex. Following the skewing ideas developed in the previous section, and taking l as in (4.4) with $0 < \lambda \leq \infty$, the skewed estimator

$$\hat{\eta}_\lambda(x) = (2\lambda + 1)^{-1}\{\lambda\,\hat{\eta}(x|x + lh) + \hat{\eta}(x|x) + \lambda\,\hat{\eta}(x|x - lh)\}$$

reduces asymptotic bias by two orders of magnitude, to h^4, and does not inflate the order of variance. Similarly, the skewed estimator defined by $\hat{\eta}_\pm(x) = \hat{\eta}(x|x \pm \kappa_2^{1/2}h)$ has asymptotic bias and asymptotic variance of sizes h^3 and $(nh)^{-1}$, respectively.

Undoubtedly, the main competitor with skewing methods applied to (approximate) local linear models (for example, (4.9) in the next section) is the fitting of (approximate) local cubics. Both methods can be shown to have the same order of asymptotic bias, and improve on mean squared error performance compared with local linear techniques. Nevertheless, we favour our skewing methods from a computational viewpoint. To see this, we note that in obtaining a local cubic estimate over a mesh of M grid points at which we wish to estimate the curve, we need to solve M *four-parameter* maximisation problems. On the other hand, skewing requires solving at most $3M$ *two-parameter* maximisation problems, and achieves the same asymptotic order of mean squared error as the local cubic method. This leads to considerable computational savings, poses less numerical challenges and confers greater numerical stability.

4.5 Regularity Conditions

The properties required of the parametric model and weight functions in the two-parameter case of Hjort and Jones' (1996) work are not stated explicitly there. However, concise conditions are needed if the technical arguments in Section 4.6 are to be clear, and so we shall be specific about them here.

Any successful candidate for g in a second-order locally parametric method has to be capable of capturing the full range of potential values of both f and its derivative. If g depends on its argument and parameters in a smooth way then this implies that, after a suitable reparametrisation, it should be approximately linear in small neighbourhoods of any given point x:

$$g(y, \theta) = \omega^{(1)} + \omega^{(2)}(y - x) + O\{(y - x)^2\} \tag{4.9}$$

as $y \to x$. Furthermore, the transformation that takes θ to $\omega = (\omega^{(1)}, \omega^{(2)})^T$ should be one-to-one and onto the whole of $(0, \infty) \times (-\infty, \infty)$. (The transformation will of course depend on x.) The differentiated forms of (4.9) must also be valid, for as many derivatives of g (with respect to y and θ, with x held fixed) that are required for other aspects of the proof. For example, we need $g'(y, \theta) = \omega^{(2)} + O(|y - x|)$ as $y \to x$.

Of course, (4.9) is satisfied by all standard two-parameter models that are used in practice in locally parametric density estimation. In particular, if g is the log-linear model employed as an example in Section 4.2, then (4.9) holds with $\omega^{(1)} = \theta^{(1)}$ and $\omega^{(2)} = \theta^{(1)}\theta^{(2)}$; and if g is the Normal model,

$$g(y, \theta) = (2\pi)^{-1/2} \left(\theta^{(2)}\right)^{-1} \exp\left\{ -\tfrac{1}{2} \left(\theta^{(2)}\right)^{-2} (y - x - \theta^{(1)})^2 \right\},$$

then (4.9) is valid with

$$\omega^{(1)} = (2\pi)^{-1/2} \left(\theta^{(2)}\right)^{-1} \exp\left\{ -\tfrac{1}{2} (\theta^{(1)}/\theta^{(2)})^2 \right\} \quad \text{and} \quad \omega^{(2)} = \omega^{(1)}\theta^{(1)} \left(\theta^{(2)}\right)^{-2}.$$

In the general formulation of locally parametric methods suggested by Hjort and Jones (1996), no explicit connection is required between the weight functions v_j and the model g. Nevertheless, their arguments implicitly ask that

$$\begin{aligned} &\text{for each } x, \text{ each of the conditions } v_j\{x, x, \theta_0(x, 0)\} \neq 0 \\ &\text{and } (\partial/\partial t)\, v_j\{x, t, \theta_0(x, 0)\}|_{t=x} \neq 0 \text{ holds for some} \\ &j = j(x) \text{ (not necessarily the same } j \text{ in both cases),} \end{aligned} \tag{4.10}$$

where, as in the previous chapter, $\theta_0(y) = \theta_0(y, h)$ is defined as the solution in θ of the equation

$$\int K_h(y - t)\, v_j(y, t, \theta) \left\{ f(t) - g(t, \theta) \right\} dt = 0 \quad \text{for} \quad j = 1, 2. \tag{4.11}$$

We assume that for each y and all sufficiently small h, $\theta_0(y)$ exists and is unique. When we intend $h = 0$ in $\theta_0(y)$, we write it as $\theta_0(y, 0)$; in all other cases, h is non-zero.

Indeed, without the second part of (4.10), $g'\{x, \hat{\theta}(x)\}$ does not approximate $f'(x)$ (see Section 4.6). Assuming that (4.9) holds and v_j is given by $v_j(x, t, \theta) = (\partial/\partial\theta^{(j)}) \log g(t, \theta)$ or $(\partial/\partial\theta^{(j)})\, g(t, \theta)$, then (4.10) is fulfilled if and only if $\omega^{(1)}$, $(\partial/\partial\theta^{(j)})\,\omega^{(1)}$ and $(\partial/\partial\theta^{(j)})\,\omega^{(2)}$ are nonzero when evaluated at $\theta = \theta_0(x, 0)$.

In nonparametric regression, local polynomial estimators typically have their *actual* bias and variance undefined. In locally parametric density estimation, however, if one fits only densities in a uniformly-bounded two-parameter class $\mathcal{G} = \{g(\cdot, \theta) : \theta \in \Theta\}$, i.e. one satisfying

$$\sup_x \sup_{\theta \in \Theta} g(x, \theta) < \infty , \tag{4.12}$$

then all the bias and variance formulae in Sections 3.4 and 4.3 (for example, (4.5) and (4.6), the latter provided that $f(x) \neq 0$) are correct as they stand, for the *actual* bias and variance. They do not represent simply the bias and variance of asymptotic distributions of $\hat{f}, \tilde{f}_\pm, \tilde{f}_\lambda$ or $\tilde{f}$.

To establish our results we need a mild additional condition on the bandwidth. It is sufficient to ask that for some $\delta > 0$ and all sufficiently large n, $h(n) \geq n^{-1+\delta}$. In company with assumptions already made, for example the condition that K be either compactly supported or the Standard Normal kernel (see Section 4.2), this may be shown to imply that for all $\epsilon, \lambda > 0$, the event $\mathcal{E} = \{|\tilde{f}_\pm(x) - f(x)| > \epsilon\}$ satisfies $P(\mathcal{E}) = O(n^{-\lambda})$ (see the similar proofs in Hall and Marron, 1997, and Cheng, Hall and Titterington, 1997). Call this result (R). Standard arguments that would be employed to establish versions of (4.5) and (4.6) when expectations are taken in asymptotic distributions, may be used to show that those results hold when, on the left-hand sides, the estimator $\tilde{f}_\pm$ (for example) is replaced by $\tilde{f}_\pm I(\widetilde{\mathcal{E}})$, where $\widetilde{\mathcal{E}}$ denotes the complement of $\mathcal{E}$ and $I(\widetilde{\mathcal{E}})$ is the indicator of $\widetilde{\mathcal{E}}$. Since, by (4.12), $0 \leq \tilde{f}_\pm \leq C$ for a finite constant C, then by (R), the mean and mean squared error of $\tilde{f}_\pm - \tilde{f}_\pm I(\widetilde{\mathcal{E}}) = \tilde{f}_\pm I(\mathcal{E})$ equal $O(n^{-\lambda})$ for all $\lambda > 0$. This allows us to make the transition from the versions of (4.5) and (4.6) for $\tilde{f}_\pm I(\widetilde{\mathcal{E}})$, to the actual formulae (4.5) and (4.6). The cases of $\tilde{f}$ or $\tilde{f}_\lambda$, rather than $\tilde{f}_\pm$, may be treated similarly.

4.6 Technical Arguments

In this section, we shall derive the bias and variance formulae stated in Section 4.3.

Theorem 4.1 *Assume that g, v_1, v_2 have properties (4.9) and (4.10) described in Section 4.5; that f, g, v_1, v_2 have four bounded derivatives with respect to each variable; that equation (4.11) has a unique solution in θ for each y; and that $h = h(n) \to$*

0 and $nh \to \infty$. Take $l = l(\lambda)$ as in (4.4). Then the following are true:

$$E(\tilde{f}_\lambda) = f + O\{h^4 + (nh)^{-1}\}, \qquad E(\tilde{f}) = f + O\{h^4 + (nh)^{-1}\},$$

$$E(\tilde{f}_\pm) = f + O\{h^3 + (nh)^{-1}\}.$$

Remark 4.1. In fact, only three bounded derivatives of f, g, v_1, v_2 are required to derive the $O\{h^3 + (nh)^{-1}\}$ bias of $\tilde{f}_\pm$.

Proof of Theorem 4.1 We denote $g^{(k)}(x,\theta)$ and $v_j^{(k)}(x,t,\theta)$ (or g with k dashes and v_j with k dashes) to mean $(\partial/\partial x)^k g(x,\theta)$ and $(\partial/\partial t)^k v_j(x,t,\theta)$ respectively. Arguing as in Hjort and Jones (1996) we may deduce that for any constant c the bias of $g\{x, \hat{\theta}(x+ch)\}$, as an estimator of $f(x)$, equals

$$g\{x, \theta_0(x+ch)\} - f(x) + O\{(nh)^{-1}\}, \tag{4.13}$$

where $\theta_0(y)$, assumed uniquely defined in a neighbourhood of x, is the solution of (4.11). Put $\delta = ch$ and Taylor-expand $\gamma(\delta) \equiv g\{x, \theta_0(x+\delta)\}$ around $\delta = 0$, as a power series in δ. The coefficient of δ in the expansion equals

$$\begin{aligned}\theta^{(1)\prime}(x)\, g_{10}\{x, \theta_0(x)\} + \theta^{(2)\prime}(x)\, g_{01}\{x, \theta_0(x)\} \\ = (\partial/\partial x)\, g\{x, \theta_0(x)\} - g'\{x, \theta_0(x)\},\end{aligned} \tag{4.14}$$

where

$$g_{jk}(y,\theta) = \{\partial^{j+k}/(\partial\theta^{(1)})^j(\partial\theta^{(2)})^k\}\, g(y,\theta).$$

To evaluate the right-hand side of (4.14) observe that, on setting $y = x$ and $\theta = \theta_0(x)$ in (4.11), and Taylor-expanding,

$$\begin{aligned}0 &= \int K_h(x-t)\, v_j\{x,t,\theta_0(x)\}\, [f(t) - g\{t,\theta_0(x)\}] \\ &= \int K(u)\, v_j\{x, x-uh, \theta_0(x)\}\, [f(x-uh) - g\{x-uh, \theta_0(x)\}]\, du \\ &= \int K(u)\, [v_j\{x,x,\theta_0(x)\} - uh\, v_j'\{x,x,\theta_0(x)\} + O(h^2)] \\ &\qquad \times \Big(f(x) - g\{x,\theta_0(x)\} - uh\, [f'(x) - g'\{x,\theta_0(x)\}] + O(h^2)\Big)\, du.\end{aligned}$$

Differentiating the right-hand side with respect to x, we obtain

$$(\partial/\partial x)\, (v_j\{x,x,\theta_0(x)\}\, [f(x) - g\{x,\theta_0(x)\}]) + O(h^2) = 0.$$

Using the product rule to evaluate the differential on the left-hand side; employing (3.12) to prove that the term $f(x) - g\{x, \theta_0(x)\}$ that forms part of the result equals $O(h^2)$; and choosing j so that $v_j\{x, x, \theta_0(x, 0)\} \neq 0$ (see (4.10)); we obtain

$$\begin{aligned} 0 &= \left[f(x) - g\{x, \theta_0(x)\}\right] (\partial/\partial x)\, v_j\{x, x, \theta_0(x)\} \\ &\qquad + v_j\{x, x, \theta_0(x)\}\left[f'(x) - (\partial/\partial x)\, g\{x, \theta_0(x)\}\right] + O(h^2) \\ &= \tfrac{1}{2}\kappa_2 h^2 \left[g''\{x, \theta_0(x)\} - f''(x)\right] (\partial/\partial x)\, v_j\{x, x, \theta_0(x)\} \\ &\qquad + v_j\{x, x, \theta_0(x)\}\left[f'(x) - (\partial/\partial x)\, g\{x, \theta_0(x)\}\right] + O(h^2)\,, \end{aligned}$$

and hence

$$f'(x) - (\partial/\partial x)\, g\{x, \theta_0(x)\} = O(h^2)\,. \tag{4.15}$$

Again Taylor-expanding the left-hand side of (4.11) with $\theta = \theta_0(x) = \theta_0(x, h)$, this time not differentiating but choosing j such that

$$(\partial/\partial t)\, v_j\{x, t, \theta_0(x, 0)\}\big|_{t=x} \neq 0 \tag{4.16}$$

(see (4.10)), we obtain

$$\begin{aligned} 0 &= \int K(u) \Big[v_j\{x, x, \theta_0(x)\} - uh\, v_j'\{x, x, \theta_0(x)\} + \tfrac{1}{2}(uh)^2 v_j''\{x, x, \theta_0(x)\} \\ &\qquad - \tfrac{1}{6}(uh)^3 v_j'''\{x, x, \theta_0(x)\} + O(h^4)\Big] \\ &\qquad \times \Big(f(x) - g\{x, \theta_0(x)\} - uh\left[f'(x) - g'\{x, \theta_0(x)\}\right] \\ &\qquad + \tfrac{1}{2}(uh)^2 \left[f''(x) - g''\{x, \theta_0(x)\}\right] \\ &\qquad - \tfrac{1}{6}(uh)^3 \left[f'''(x) - g'''\{x, \theta_0(x)\}\right] + O(h^4)\Big)\, du \\ &= v_j\{x, x, \theta_0(x)\}\left[f(x) - g\{x, \theta_0(x)\}\right] \\ &\qquad + \tfrac{1}{2}\kappa_2 h^2 \Big(v_j\{x, x, \theta_0(x)\}\left[f''(x) - g''\{x, \theta_0(x)\}\right] \\ &\qquad + v_j''\{x, x, \theta_0(x)\}\left[f(x) - g\{x, \theta_0(x)\}\right] \\ &\qquad + 2\, v_j'\{x, x, \theta_0(x)\}\left[f'(x) - g'\{x, \theta_0(x)\}\right]\Big) + O(h^4)\,, \end{aligned}$$

noting that K is symmetric and thus terms of order h^3 vanish. Using (3.12) we deduce that the left-hand side equals

$$\kappa_2 h^2\, v_j'\{x, x, \theta_0(x)\}\left[f'(x) - g'\{x, \theta_0(x)\}\right] + O(h^4)\,,$$

whence it follows from (4.16) that

$$f'(x) - g'\{x, \theta_0(x)\} = O(h^2)\,. \tag{4.17}$$

Combining (4.15) and (4.17) we see that the right-hand side of (4.14) equals $O(h^2)$. Hence, the term in δ in the Taylor expansion of $\gamma(\delta)$ is of size $O(\delta h^2) = O(h^3)$.

Next we deal with the coefficient of $\frac{1}{2}\delta^2$, which may be shown by an analogue of the argument leading to (4.14) to equal

$$(\partial/\partial x)^2 g\{x, \theta_0(x)\} + g''\{x, \theta_0(x)\} - 2\,(\partial/\partial x)\, g'\{x, \theta_0(x)\}\,. \tag{4.18}$$

Formally differentiating (3.12) we deduce that $(\partial/\partial x)^2[g\{x, \theta_0(x)\} - f(x)] = O(h^2)$. This result may be obtained rigorously by making minor modifications to arguments of Hjort and Jones (1996). Refining the argument leading to (4.17) we may identify the right-hand side and show that, after one differentiation, it is still of order $O(h^2)$. Therefore, $f''(x) - (\partial/\partial x)\, g'\{x, \theta_0(x)\} = O(h^2)$. Combining the last two results we see that the quantity at (4.18) equals $g''\{x, \theta_0(x)\} - f''(x) + O(h^2)$. From this formula for the coefficient of $\frac{1}{2}\delta^2$ in the Taylor expansion of $\gamma(\delta)$, and from the result in the previous paragraph for the coefficient of δ, we deduce that

$$g\{x, \theta_0(x+\delta)\} - g\{x, \theta_0(x)\} = \tfrac{1}{2}\delta^2\,[g''\{x, \theta_0(x)\} - f''(x)] + O(h^3)\,. \tag{4.19}$$

Using (3.12) and (4.19) we find that the quantity at (4.13) (equal to the bias of $g\{x, \hat\theta(x+ch)\}$) equals

$$\tfrac{1}{2}\,h^2(\kappa_2 - c^2)\,[f''(x) - g''\{x, \theta_0(x)\}] + O\{h^3 + (nh)^{-1}\}\,.$$

Since $\tilde f_\pm$ is defined by taking $c = \pm\kappa_2^{1/2}$ in $g\{x, \hat\theta(x+ch)\}$ then, for either choice of the $+$ and $-$ signs, its bias equals simply $O\{h^3 + (nh)^{-1}\}$.

Appealing to symmetry properties when evaluating Taylor expansions, and employing a similar but longer argument than that leading to (4.19), it may be proved that

$$\begin{aligned} &g\{x, \theta_0(x+ch)\} + g\{x, \theta_0(x-ch)\} - 2\,f(x) \\ &\qquad = h^2(\kappa_2 - c^2)\,[f''(x) - g''\{x, \theta_0(x)\}] + O(h^4)\,. \end{aligned}$$

From this formula and (3.12) we deduce that

$$(2\lambda+1)^{-1}\Big(\lambda\,[g\{x, \theta_0(x+ch)\} + g\{x, \theta_0(x-ch)\}] + g\{x, \theta_0(x)\}\Big) - f(x)$$

$$= \frac{h^2}{2}(2\lambda+1)^{-1}\{(2\lambda+1)\,\kappa_2 - 2\lambda c^2\}\,[f''(x) - g''\{x,\theta_0(x)\}] + O(h^4)\,. \tag{4.20}$$

The left-hand side equals the bias of $\hat{f}_\lambda$, up to terms of order $(nh)^{-1}$. Taking $c = l$, where l is defined by (4.4), the right-hand side of (4.20) equals $O(h^4)$. Hence, $E(\tilde{f}_\lambda) - f = O\{h^4 + (nh)^{-1}\}$. Similarly we may prove that the bias of $\tilde{f} = \tilde{f}_\infty$ equals $O\{h^4 + (nh)^{-1}\}$.

Theorem 4.2 *Assume that g, v_1, v_2 satisfy the conditions in Theorem 4.1, that (4.9) and its differentiated form holds, and g is uniformly-bounded (see (4.12)). Then,*

$$\mathrm{var}\,(\tilde{f}_\pm) \sim (nh)^{-1}\,(\kappa_1 + \kappa_2^{-1}\kappa_3)\,f\,, \quad \mathrm{var}\,(\tilde{f}_\lambda) \sim (nh)^{-1}\,V(\lambda)\,f\,,$$

as $h \to 0$ and $n \to \infty$ such that $nh \to \infty$, where $\kappa_1 = \int K^2, \kappa_2 = \int t^2\,K(t)\,dt, \kappa_3 = \int t^2 K(t)^2\,dt$ and $V(\lambda)$ is given by (4.7).

Proof of Theorem 4.2 Without loss of generality, we assume the original parametrisation of g was $\theta = \omega$. Following the argument in Section 4.2 of Hjort and Jones (1996), the variance of $\tilde{f}_\pm$ is seen to be asymptotic to $(nh)^{-1}\tau(K)^2 f(x)$, where, in place of Hjort and Jones' formula for $\tau(K)^2$, one has

$$\tau(K)^2 = w_1^T M_1^{-1} M_2 M_1^{-1} w_1\,,$$

with $w_1^T = (1 + o(1), ch + o(h))$, $M_1 = \mathrm{diag}\,(1, h^2\kappa_2)$, $M_2 = \mathrm{diag}\,(\kappa_1, h^2\kappa_3)$ and $c = \pm\kappa_2^{1/2}$. It follows that $\tau(K)^2 \sim \kappa_1 + c^2\kappa_2^{-2}\kappa_3 = \kappa_1 + \kappa_2^{-1}\kappa_3$, as had to be proved.

Formula for the variances of $\tilde{f}_\lambda$ and $\tilde{f}$ may be derived by similar but more elaborate arguments. To calculate the variance for $\tilde{f}_\lambda$, we need to obtain the leading terms in $\mathrm{var}\,\{\hat{f}(x|\,x \pm lh)\}$, $\mathrm{cov}\,\{\hat{f}(x|x), \hat{f}(x|\,x \pm lh)\}$ and $\mathrm{cov}\,\{\hat{f}(x|\,x - lh), \hat{f}(x|\,x + lh)\}$. The first term is shown in the previous paragraph to be $(nh)^{-1}\,\{\kappa_1 + (l\kappa_2^{-1})^2\kappa_3\}\,f(x)$. By similar arguments, $\mathrm{cov}\,\{\hat{f}(x|x), \hat{f}(x|\,x \pm lh)\}$ may be seen to be asymptotic to $(nh)^{-1}\,\nu(K)^2_\pm\,f(x)$, where $\nu(K)^2_\pm = w_2^T M_1^{-1} M_3 M_1^{-1} w_3$ with $w_2^T = (1 + o(1), lh + o(h))$, $w_3^T = (1 + o(1), o(1))$ and

$$M_3 = \begin{pmatrix} \int K(u)\,K(u-l)\,du & h\int u\,K(u)\,K(u-l)\,du \\ h\int (u+l)\,K(u)\,K(u-l)\,du & h^2\int u(u+l)\,K(u)\,K(u-l)\,du \end{pmatrix}.$$

Carrying out the required matrix algebra, we obtain

$$\nu(K)^2_\pm = \int K(u)\,K(u \pm l)\,du \pm l\kappa^{-1}\int (u \pm l)\,K(u)\,K(u \pm l)\,du\,.$$

The first-order term of $\operatorname{cov}\{\hat{f}(x|\,x-lh), \hat{f}(x|\,x+lh)\}$ may be obtained in a similar fashion, which equals $(nh)^{-1}\,\zeta(K)^2\,f(x)$, where $\zeta(K)^2 = w_2^T M_1^{-1} M_4 M_1^{-1} w_4$ with $\omega_4^T = (1+o(1), -lh+o(h))$ and

$$M_4 = \begin{pmatrix} \int K(u+l)\,K(u-l)\,du & h\int(-l+u)\,K(u+l)\,K(u-l)\,du \\ h\int(l+u)\,K(u+l)\,K(u-l)\,du & h^2\int(u^2-l^2)\,K(u+l)\,K(u-l)\,du \end{pmatrix}.$$

Again, performing the required matrix multiplication gives

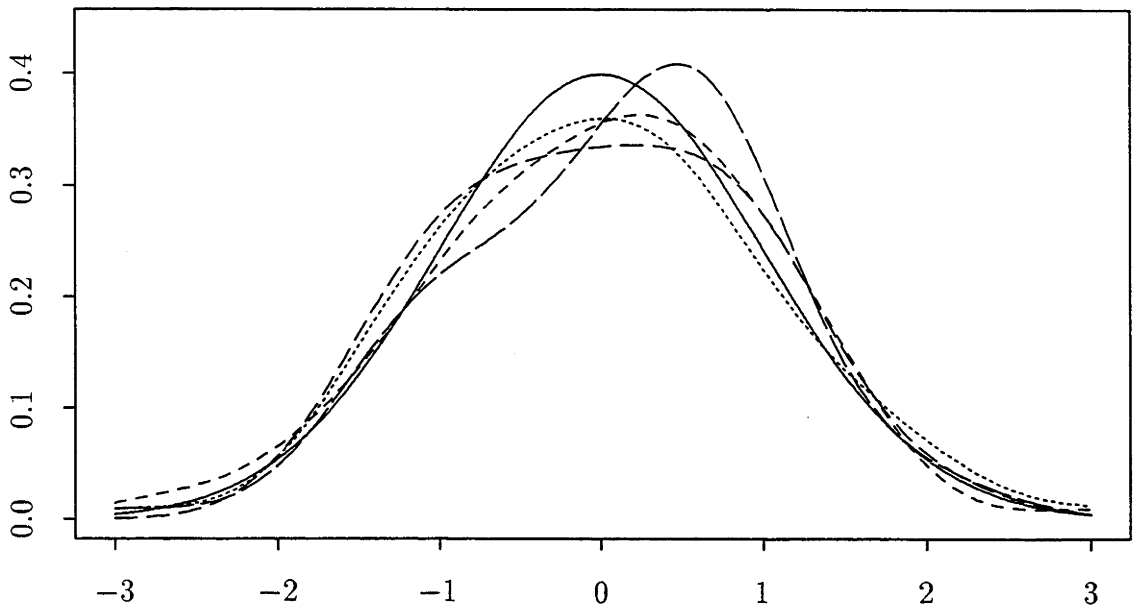
$$\zeta(K)^2 = \kappa_2^{-1}(\kappa_2 + 2l^2)\int K(u+l)\,K(u-l)\,du \\ -(\kappa_2^{-1}l)^2\int(u^2-l^2)\,K(u+l)\,K(u-l)\,du\,.$$

Combining the above results yields the required variance expressions for various versions of skewed density estimators.

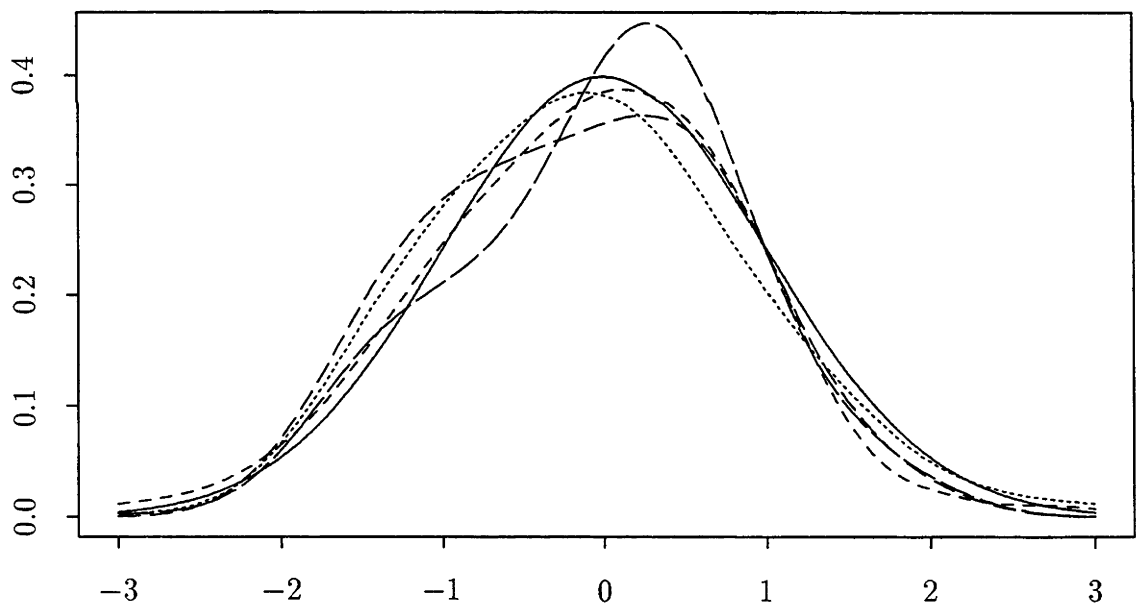
4.7 Numerical Properties

We summarise here a simulation study which examines finite-sample properties of various skewed estimators. We chose 5 densities, $f_1, \ldots, f_5$, namely "Gaussian", "skewed unimodal", "bimodal", "separated bimodal" and "asymmetric bimodal" as described by Marron and Wand (1992). We used sample sizes $n = 50, 100, 200, 500$ and 1000, although only results for $n = 100$ and for the "Gaussian" and "skewed unimodal" densities will be discussed in detail. Results for other values of n and other densities are similar, and their mean integrated squared error (MISE) performances will be summarised in Table 4.1. The kernel methods with which we chose to compare skewing were a standard second-order kernel estimator $\bar{f}$, using the Standard Normal kernel ϕ, and a fourth-order kernel estimator $\bar{f}_{(4)}$, based on the kernel $K(x) = \frac{1}{2}(3-x^2)\,\phi(x)$. See the monograph by Wand and Jones (1995, Chapter 2) on higher-order kernels in density estimation.

The locally parametric density estimators employed here were constructed using the local log-linear parametric model $g(y,\theta) = \theta^{(1)}\exp\{(y-x)\theta^{(2)}\}$, because of its popularity (e.g. Hjort and Jones, 1996; Loader, 1996), its simplicity (e.g. the availability of the closed-form estimator at (4.2)), and the central position occupied by local linear methods in contemporary curve estimation. However, other choices of g can better capture local features of a density, leading to improved performance

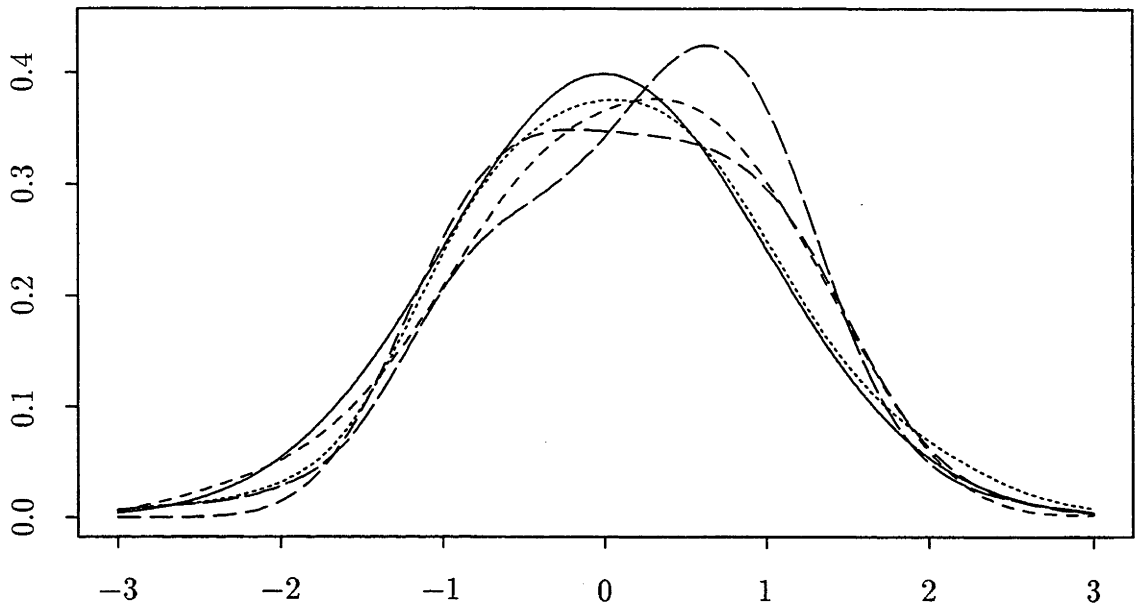


(a)

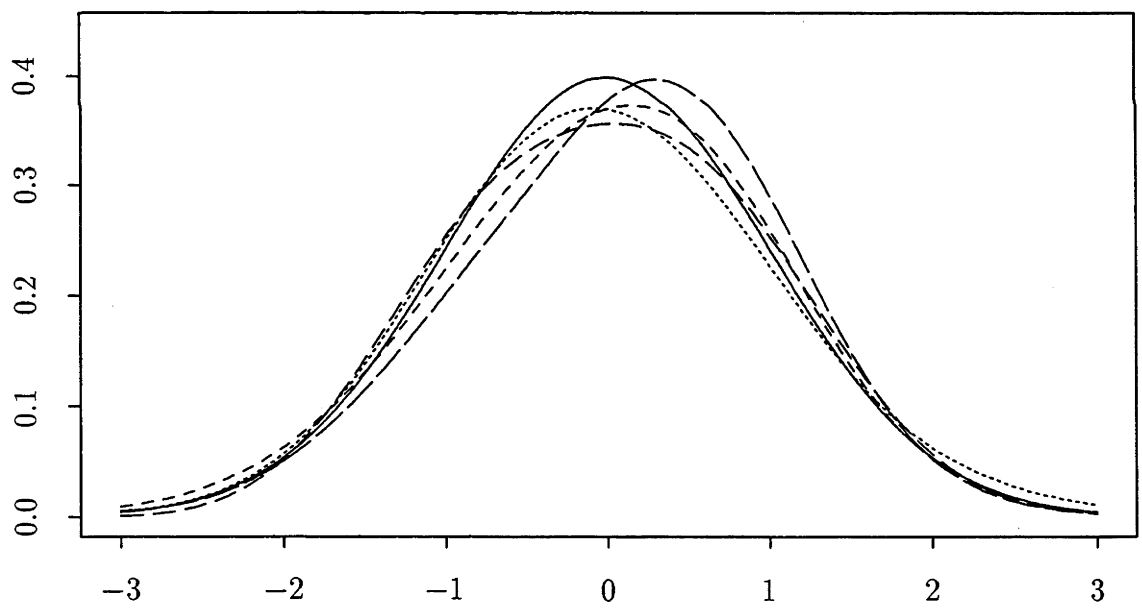


(b)

Figure 4.1(i): Density estimates for the symmetric unimodal density f_1 with sample size $n = 100$. Panels (a) and (b) depict the estimates $\bar{f}$ and $\tilde{f}_+$ respectively. The solid line represents the true density f, and broken lines show its estimates. The same four data sets are used in all panels, with their respective line types kept the same. The respective bandwidths employed for $\bar{f}$ and $\tilde{f}_+$ were 0.450 and 0.539.

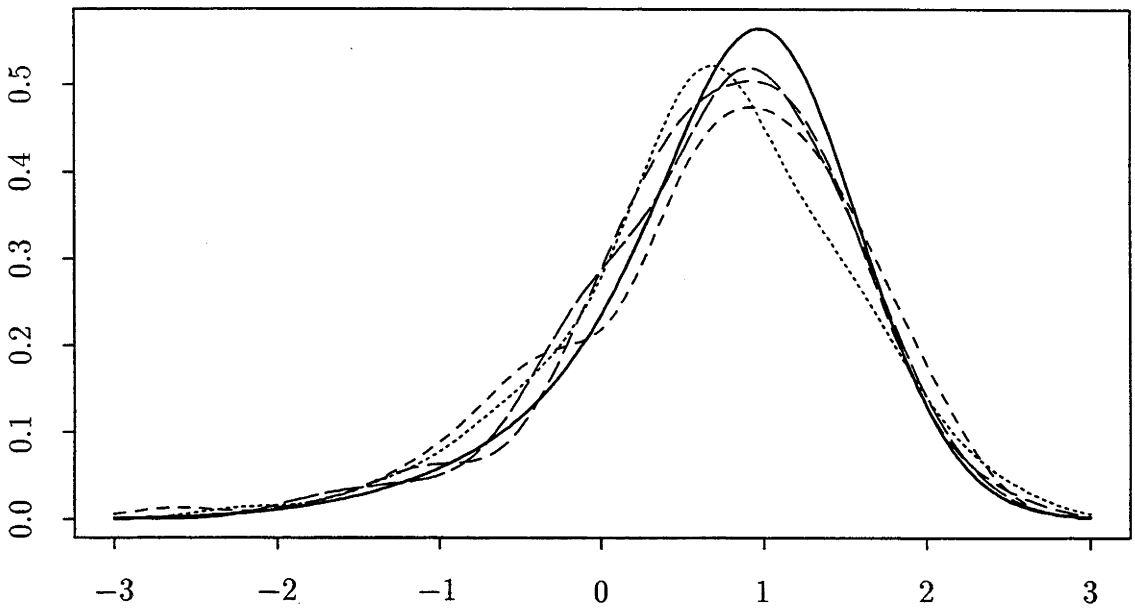


(c)

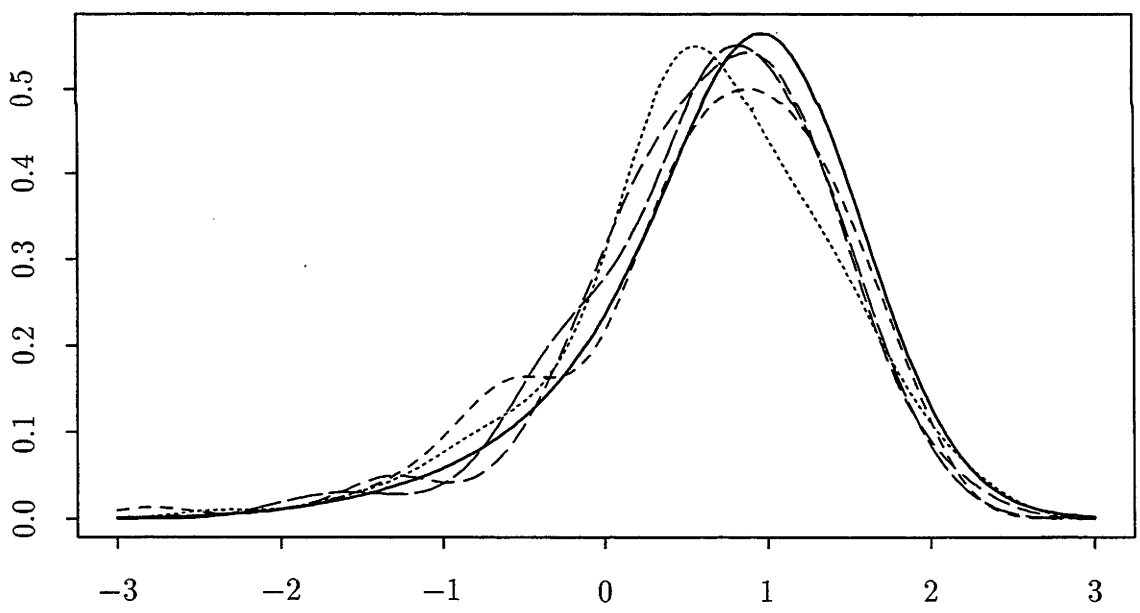


(d)

Figure 4.1(ii): Density estimates for the symmetric unimodal density f_1 with sample size $n = 100$. Panels (c) and (d) depict the estimates $\tilde{f}_-$ and $\tilde{f}$, respectively. The solid line represents the true density f, and broken lines show its estimates. The same four data sets are used in all panels, with their respective line types kept the same. The respective bandwidths employed for $\tilde{f}_-$ and $\tilde{f}$ were 0.541 and 0.645.

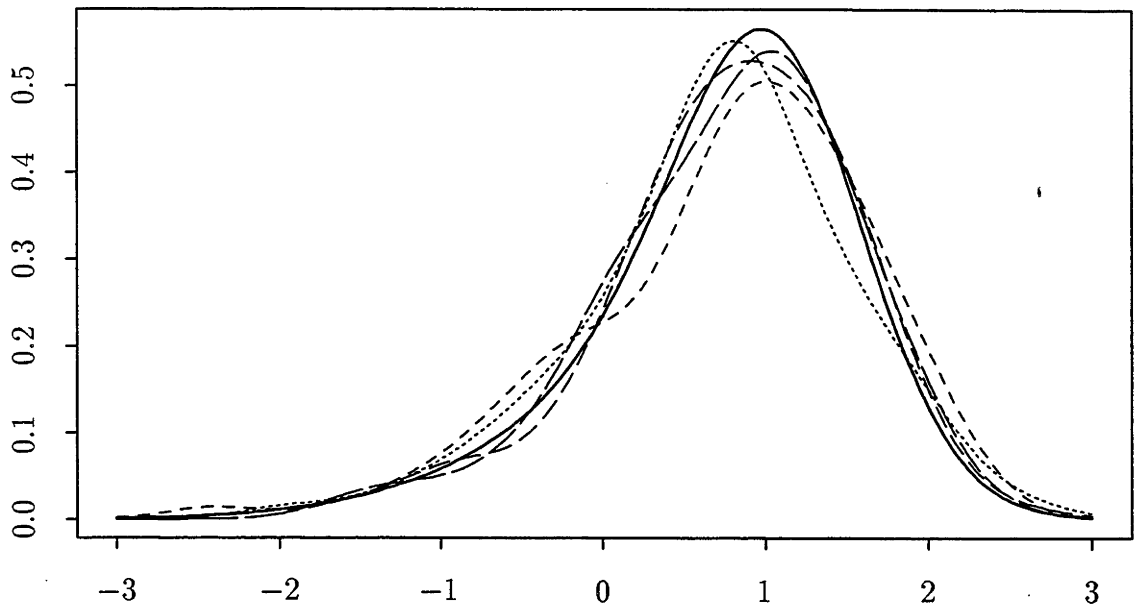


(a)

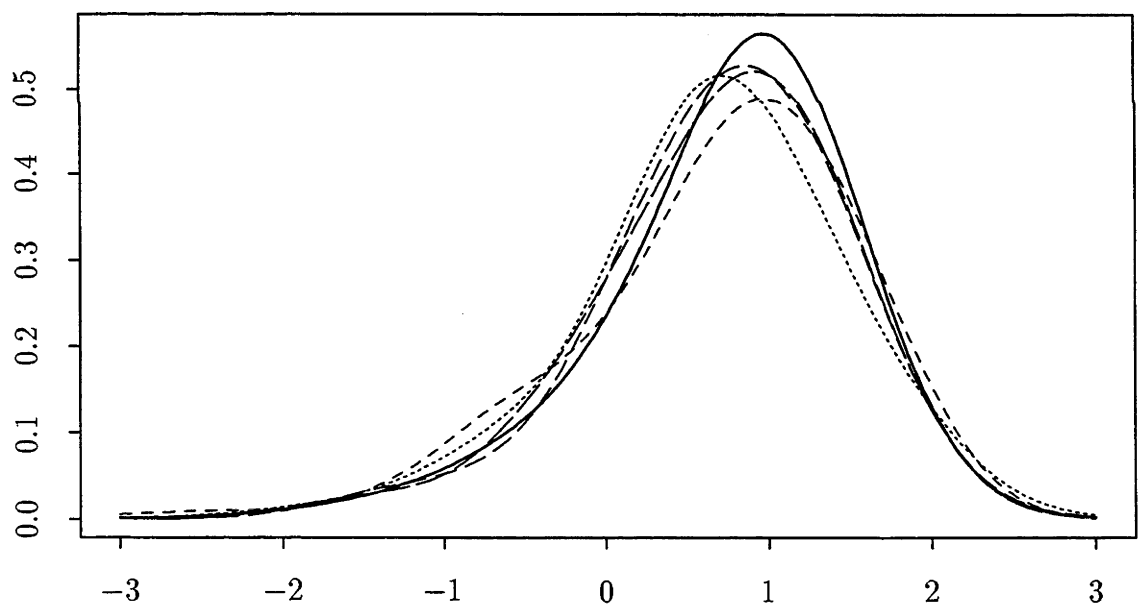


(b)

Figure 4.2(i): Density estimates for the skewed unimodal density f_2 with sample size $n = 100$. Panels (a) and (b) depict the estimates $\bar{f}$ and $\tilde{f}_+$ respectively. The solid line represents the true density f, and broken lines show its estimates. The same four data sets are used in all panels, with their respective line types kept the same. The respective bandwidths employed for $\bar{f}$ and $\tilde{f}_+$ were 0.306 and 0.358.



(c)



(d)

Figure 4.2(ii): Density estimates for the skewed unimodal density f_2 with sample size $n = 100$. Panels (c) and (d) depict the estimates $\tilde{f}_-$ and $\tilde{f}$, respectively. The solid line represents the true density f, and broken lines show its estimates. The same four data sets are used in all panels, with their respective line types kept the same. The respective bandwidths employed for $\tilde{f}_-$ and $\tilde{f}$ were 0.382 and 0.429.

of the locally parametric estimators $\hat{f}_0(x) = \hat{f}(x|x)$, $\tilde{f}_\pm(x) = \hat{f}(x|x \pm h)$ and $\tilde{f} = \frac{1}{2}(\tilde{f}_+ + \tilde{f}_-)$.

Four typical realisations of $\bar{f}$, $\tilde{f}_\pm$ and $\tilde{f}$ in the case of estimating f_1 and f_2 are displayed in Figures 4.1 and 4.2 respectively, with $n = 100$. For each estimator, the bandwidth used was the minimiser of its mean integrated squared error, approximated using the simulation study described in the next paragraph. The three locally parametric estimators $\tilde{f}_\pm$ and $\tilde{f}$ may be interpreted as attempts at improving the shape of the peaks. The "sharpness" of the peak is captured better by $\tilde{f}_\pm$, although all density estimates constructed using $\tilde{f}_+$ are shifted to the left, and shifted to the right in the case of $\tilde{f}_-$. The estimator $\tilde{f}$ approximates the peak more accurately than do either $\tilde{f}_+$ or $\tilde{f}_-$, and provides better approximations in the tails. These are reflected in the smaller pointwise mean squared errors (PMSE) of $\tilde{f}$ at both the peak and the tails in estimating the Gaussian and skewed unimodal densities. For the sake of brevity, we have not included plots of PMSE here.

To calculate the MISE curves we used a grid of bandwidths consisting of 51 logarithmically equally-spaced points in the interval $[0.1, 1.0]$. Each MISE curve was obtained by averaging 1000 replications of integrated squared error (ISE) curves. For each bandwidth h in the grid, we calculated the pointwise squared errors of the estimates at 201 equally-spaced points on the interval $[-3, 3]$. The trapezoidal rule was employed to evaluate ISE. The MISE curves for $n = 100$ and for the densities f_1 and f_2 are depicted in Figures 4.3 and 4.4 respectively. For the sake of clarity, only bandwidths in the interval $[0.15, 1.0]$ are displayed. Vertical lines are drawn through the minimisers of the MISE curves, and have the same line types as the respective curves.

For the symmetric density f_1, the estimator $\tilde{f}$ performs better than $\tilde{f}_\pm$ throughout the range of bandwidths considered. In the case of small bandwidths, the MISE curves for the standard kernel estimator $\bar{f}$ and the standard locally parametric estimator $\hat{f}_0$ are almost identical, whereas discrepancies are noticeable for large h. This is to be expected since, as mentioned by Hjort and Jones (1996), for small to moderate h the locally parametric estimator utilises primarily local properties of the model g, and hence the estimation method is essentially nonparametric. As h increases the method becomes more parametric, and the difference between MISE curves is best explained by errors in approximating the true density by the model. Note, however, that the minimum MISE's for $\bar{f}$ and $\hat{f}_0$ are approximately equal in

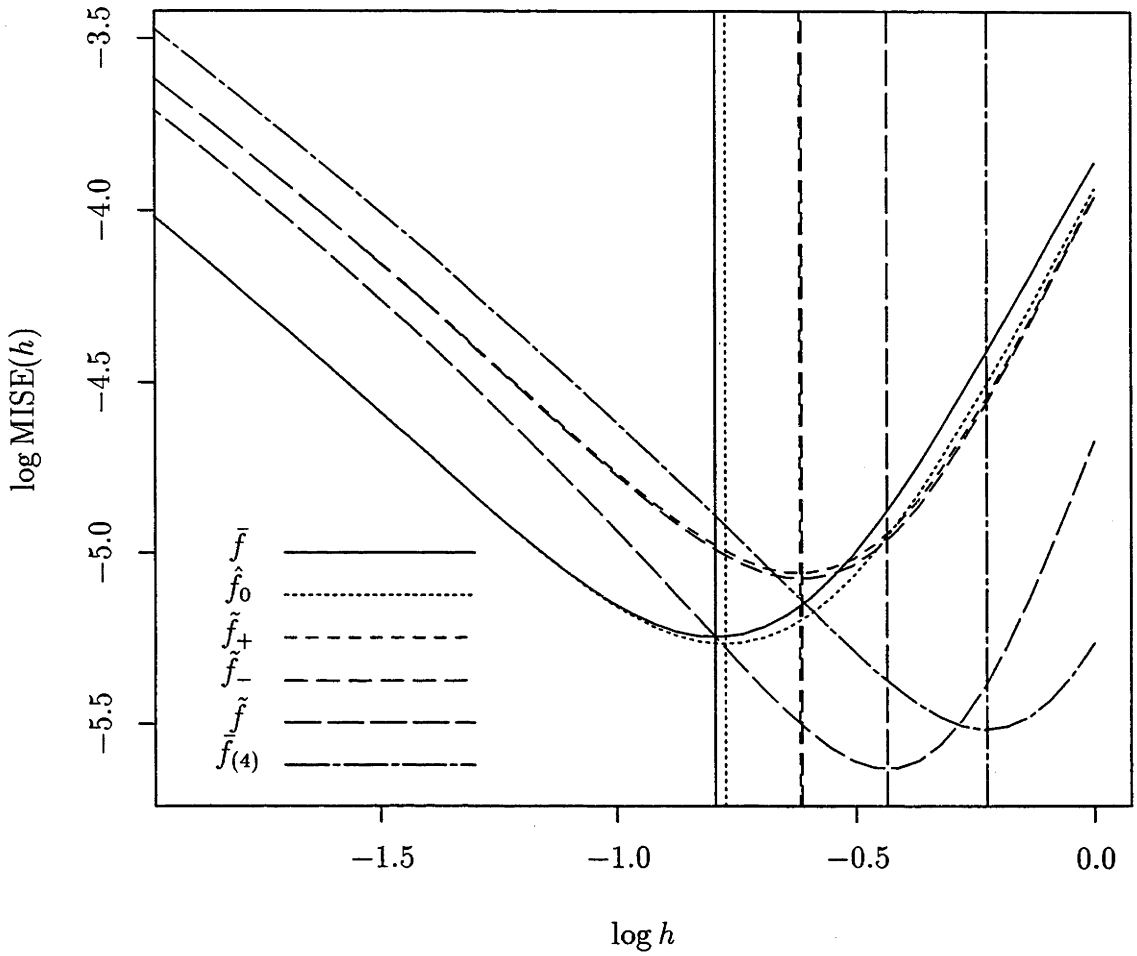


Figure 4.3: Comparison of MISE curves for the Gaussian density f_1 with sample size $n = 100$. The figure is plotted on a log-log scale, for the sake of clarity. The line types in the legend correspond to the estimators $\bar{f}$, $\hat{f}_0$, $\tilde{f}_+$, $\tilde{f}_-$, $\tilde{f}$ and $\bar{f}_{(4)}$ respectively.

all our simulations, as illustrated in Table 4.1.

The performance of $\tilde{f}_\pm$ improves on that of $\hat{f}_0$ for large n, although not necessarily for smaller sample sizes. This is illustrated in Figures 4.3 and 4.4, where the minimum MISE for $\hat{f}_0$ is seen to be less than that for $\tilde{f}_\pm$ in the case $n = 100$. From the asymptotic theory, $\tilde{f}_\pm$ should outperform $\bar{f}$ when the sample size n is large enough. Our simulation results indicate that this will be the case when n is larger than 1000, as shown by the increasing efficiency as a function of n in Table 4.1. On

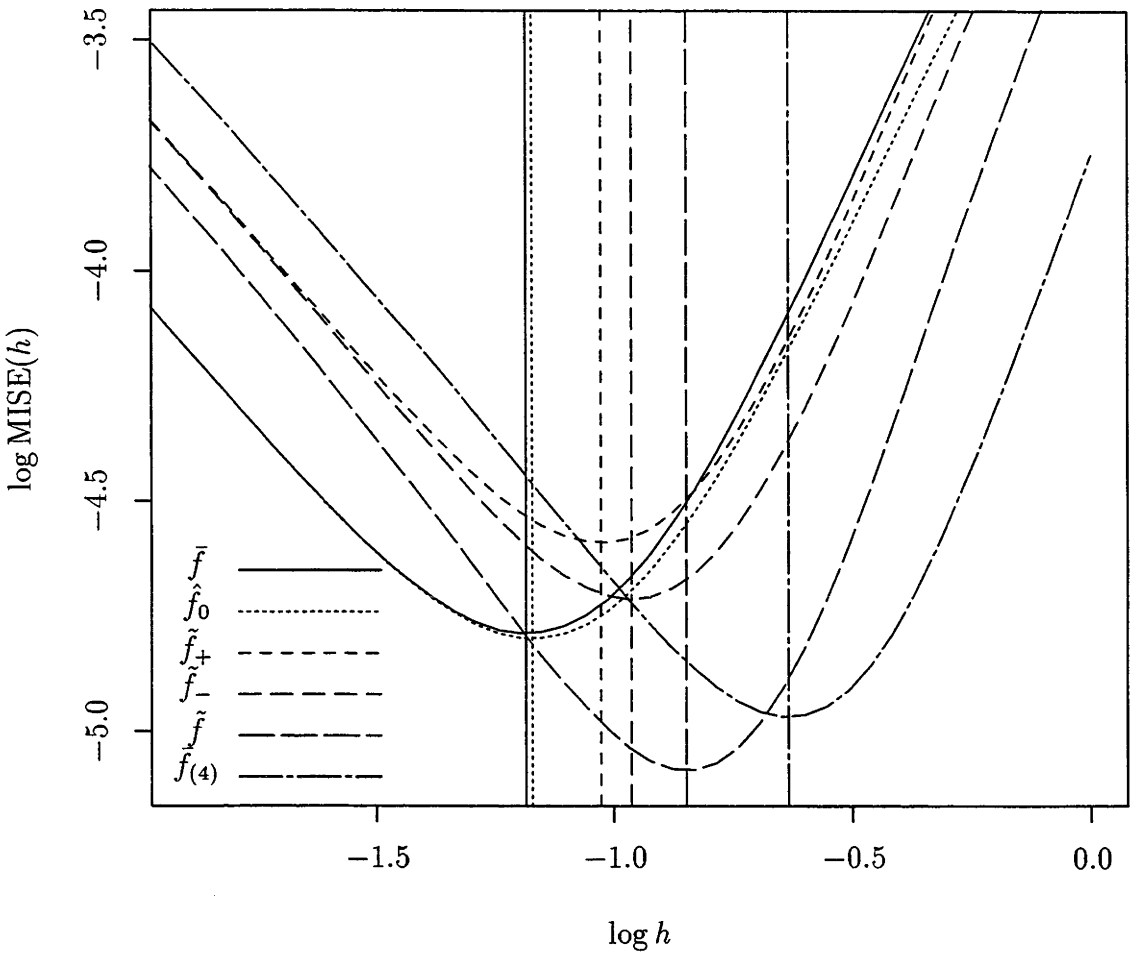


Figure 4.4: Comparison of MISE curves for the skewed unimodal density f_2 with sample size $n = 100$. Again, the figure is plotted on a log-log scale. The line types in the legend correspond to the estimators $\bar{f}$, $\hat{f}_0$, $\tilde{f}_+$, $\tilde{f}_-$, $\tilde{f}$ and $\bar{f}_{(4)}$, respectively.

the other hand, $\tilde{f}$ outperforms all its competitors in almost all our simulations, even for sample sizes as small as $n = 100$. In particular, $\tilde{f}$ achieves greater efficiency than $\bar{f}_{(4)}$ in all cases, even though both estimators have biases of size h^4.

The estimators $\hat{f}_0$, $\tilde{f}_\pm$ and $\tilde{f}$ in general do not integrate to 1, and normalising the estimators by dividing by their respective integrals may improve finite-sample performance. For example, in the case of Standard Normal data, the improvement in mean integrated squared error of $\tilde{f}$ is by 10% when $n = 100$, with smaller increases for other densities in our simulation study. Similar results were obtained by Jones,

Linton and Nielsen (1995) and Jones and Signorini (1997). Our simulation results also indicate that normalisation has minimal effects on $\hat{f}_0$ and $\tilde{f}_{\pm}$.

Estimator	Sample Size n				
	50	100	200	500	1000
			Unimodal		
$\hat{f}_0$	1.013	1.019	1.018	1.014	1.011
$\tilde{f}_+$	0.759	0.820	0.852	0.916	0.958
$\tilde{f}_-$	0.782	0.834	0.869	0.939	0.963
$\tilde{f}$	1.378	1.455	1.524	1.600	1.675
$\bar{f}_{(4)}$	1.199	1.296	1.329	1.429	1.492
			Skewed Unimodal		
$\hat{f}_0$	1.014	1.010	1.009	1.005	1.004
$\tilde{f}_+$	0.745	0.803	0.827	0.894	0.938
$\tilde{f}_-$	0.860	0.908	0.914	0.958	0.974
$\tilde{f}$	1.270	1.320	1.373	1.445	1.518
$\bar{f}_{(4)}$	1.115	1.174	1.237	1.307	1.384
			Bimodal		
$\hat{f}_0$	1.019	1.013	1.009	1.007	1.004
$\tilde{f}_+$	0.736	0.753	0.825	0.891	0.904
$\tilde{f}_-$	0.733	0.792	0.822	0.865	0.919
$\tilde{f}$	0.989	1.057	1.137	1.230	1.291
$\bar{f}_{(4)}$	0.950	1.004	1.076	1.164	1.224
			Separated Bimodal		
$\hat{f}_0$	1.012	1.009	1.006	1.005	1.003
$\tilde{f}_+$	0.790	0.822	0.863	0.910	0.937
$\tilde{f}_-$	0.777	0.813	0.861	0.902	0.950
$\tilde{f}$	1.186	1.247	1.315	1.411	1.479
$\bar{f}_{(4)}$	1.091	1.140	1.120	1.283	1.353
			Asymmetric Bimodal		
$\hat{f}_0$	1.018	1.011	1.006	1.004	1.002
$\tilde{f}_+$	0.709	0.756	0.786	0.827	0.836
$\tilde{f}_-$	0.783	0.803	0.836	0.865	0.899
$\tilde{f}$	0.940	0.982	1.028	1.101	1.136
$\bar{f}_{(4)}$	0.926	0.959	1.006	1.068	1.110

Table 4.1: *Relative "efficiencies" (i.e. ratios of the minimum MISE values) of* $\hat{f}_0$, $\tilde{f}_+$, $\tilde{f}_-$, $\tilde{f}$ *and* $\bar{f}_{(4)}$ *relative to the standard kernel estimator* $\bar{f}$, *calculated for five of the fifteen Gaussian mixture densities of Marron and Wand (1992).*

Chapter 5

Estimating Intensity Surfaces and Correlation Dimensions

5.1 Introduction

The main theme of earlier chapters of this thesis was nonparametric curve estimation, where we introduced new bias-reduction techniques in regression analysis and density estimation. The material in the rest of the thesis is related to nonparametric surface estimation, where the surface represents the intensity of a point process in the plane. The surface contains poles – that is, places where it is infinite – and the strength of a pole is related to the *correlation dimension* of the point process in the neighbourhood of the pole. We shall first review existing methods for estimating correlation dimensions of point patterns, and then go on to investigate properties of poles in intensity functions in the next chapter. Technical terms will be defined in subsequent sections.

For statisticians working with scientists in seismology, analyses of earthquake patterns using fractal models are gaining increasing popularity (Hirata and Imoto, 1991; Eneva, 1996; Harte, 1996; Vere-Jones *et al.*, 1997). It is argued that underlying "self-organising" processes control energy release from tectonic processes to microfractures. *Fractal dimensions* are estimated for different earthquake catalogues and are used to compare the physical characteristics of different earthquake regions. If the data consist only of latitudes, longitudes and depths, i.e. the spatial aspects of the events, then the dimension may be able to provide information about the underlying geometry of the fault systems. An example is given in Figures 5.1

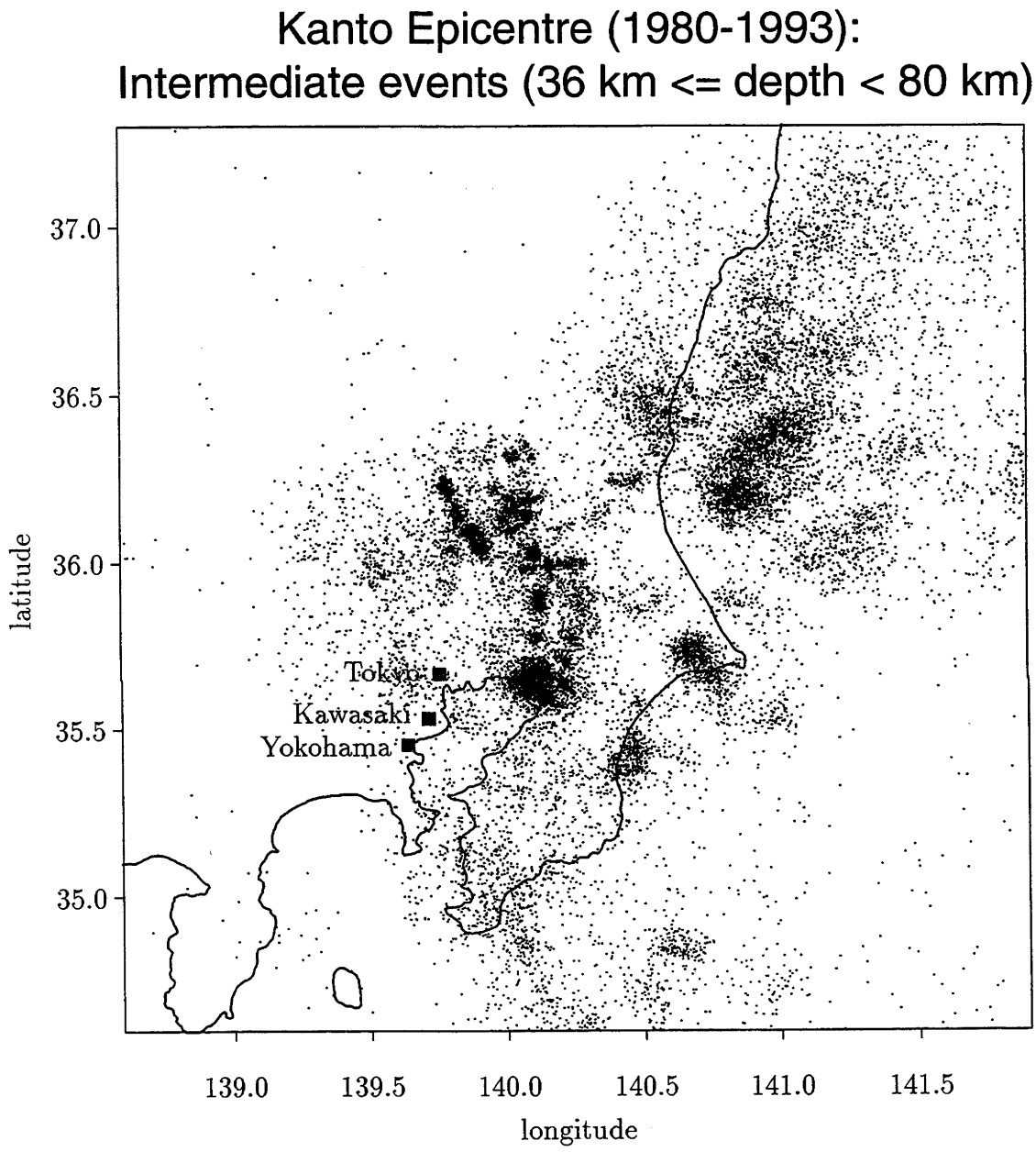


Figure 5.1: Kanto epicentres for 36km ≤ depth < 80km. The longitude ranges from 138.6° to 141.9° and the latitude ranges from 34.6° to 37.3°. All events have magnitude at least 2.0.

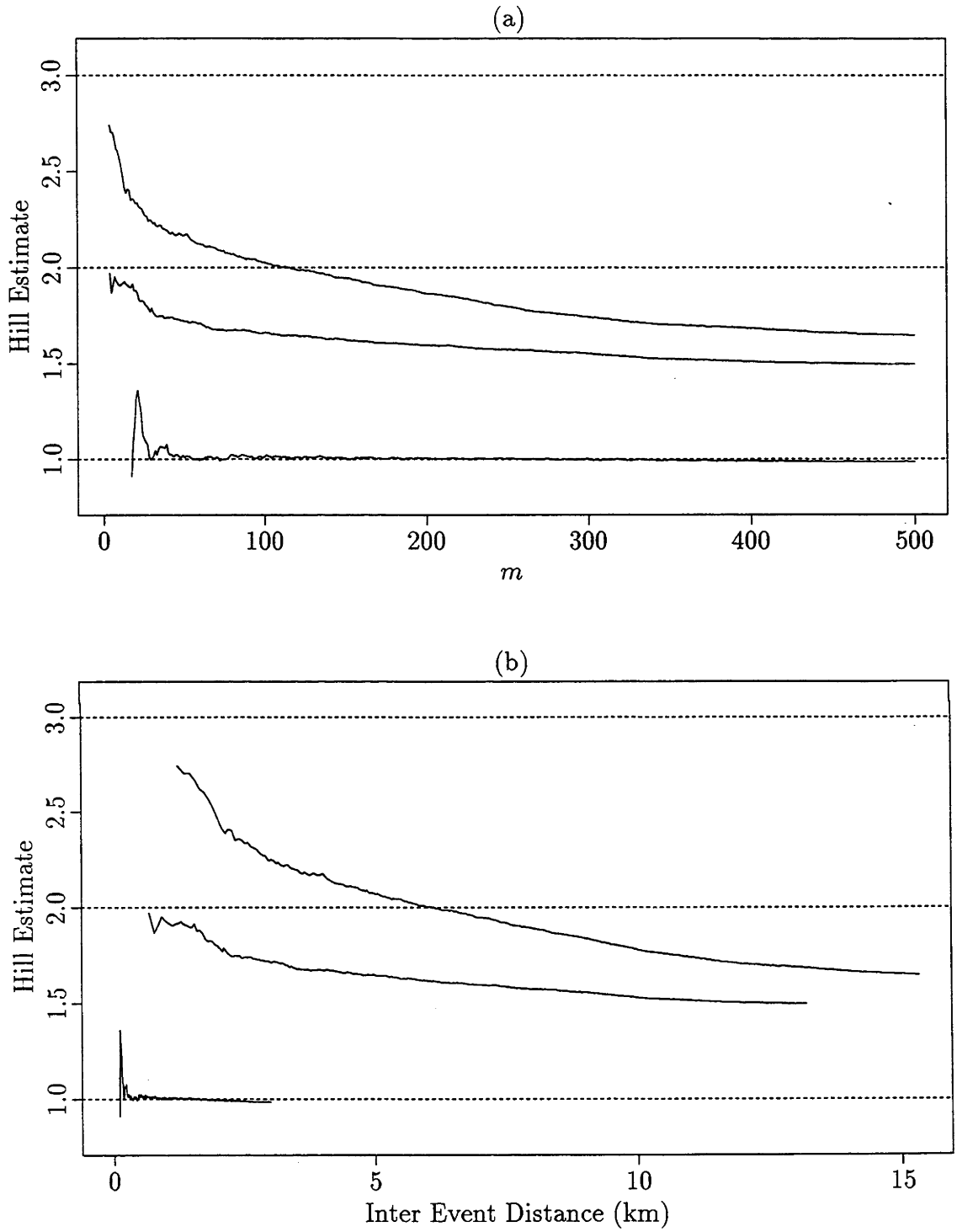


Figure 5.2: Dimension estimates for Kanto intermediate events. Panel (a) depicts the Hill estimates, employing different number m of order statistics. The inter-event distance in panel (b) is calculated as the average over all $B = 100$ bootstraps of the m-th order statistic.

and 5.2, where the correlation dimension is calculated for the Kanto region. Figure 5.1 depicts epicentres of events of magnitude at least 2.0 on the Richter scale, occurring at depths between 36 km and 80 km during the period 1980-1993. Figure 5.2 shows the dimension estimates using the method of Hill (1975). The bottom, middle and top lines represent the dimension estimates using longitudes, epicentres and hypocentres respectively, with respective estimates of 1.0, 1.5 and 1.7. The estimation procedure will be explained in more detail in Section 5.3.

Vere-Jones (1996) pointed out that there is ambiguity as to the interpretation of such fractal dimensions. He suggested two possible scenarios, and in the case of earthquake data, it may be more appropriate to consider the estimates as the fractal dimension associated with the spatial intensity of the process. Nevertheless, this interpretation does not appear to give a clear intuition of the actual features of the process. Moreover, there are numerous definitions of a fractal dimension. As noted by Hentschel and Procaccia (1983), different dimensions are not all equivalent. Cutler (1991) gave a more careful discussion of different definitions of dimension and the conditions under which they are equal.

We shall only concentrate on the problem of estimating correlation dimension in this chapter. In the seismological context, Vere-Jones *et al.* (1997) argued that the implication of including time when estimating dimension in addition to the spatial coordinates is less clear. We shall ignore the time aspect of the data, and focus on the "static" problem of dimension estimation. Correlation dimension will be defined and various estimation methods will be discussed, together with some of their statistical properties. The methods we shall review include works by Grassberger and Procaccia (1983a, 1983b, 1983c), Takens (1985), Smith (1992) and Mikosch and Wang (1995). This chapter does not attempt to give an exhaustive review of the different aspects of fractal dimension estimation, and we shall concentrate only on those aspects which are most relevant to our work.

5.2 Correlation Dimension and the Grassberger-Procaccia Procedure

Let $\mathcal{A}$ be a subset of the d-dimensional Euclidean space $\mathbb{R}^d$, where our observations $X_1, \ldots, X_N$ lie. Let μ be a probability measure on $\mathcal{A}$, and let $\mathcal{S}(x, \epsilon)$ denote a sphere of radius ϵ centred on x. For any integer q, the *correlation integral*, if it exists, is

defined as

$$\bar{C}_q(\mu, \epsilon) = \int_{\mathcal{A}} \mu\{\mathcal{S}(x, \epsilon)\}^{q-1} \mu(dx)\,, \tag{5.1}$$

although it is more generally referred to in the case $q = 2$. Note that if $q = 2$, $\bar{C}_q(\mu, \epsilon)$ may be expressed as

$$C(\epsilon) = P(\|U - V\| \leq \epsilon)\,,$$

where U, V are independent and identically distributed with respect to the probability measure μ, and $\|\cdot\|$ is an appropriate norm in $\mathbb{R}^d$. Thus, the correlation integral (for $q = 2$) can be thought of as the distribution function of $\|U - V\|$. The *correlation dimension* or *correlation exponent* (with embedding dimension d) is defined as

$$\alpha = \lim_{\epsilon \to 0} \frac{\log C(\epsilon)}{\log \epsilon}\,, \tag{5.2}$$

if the limit exists. It can be easily shown that the correlation dimension exists if and only if

$$C(\epsilon) = g(\epsilon)\, \epsilon^{\alpha}\,, \tag{5.3}$$

where g is a positive function such that $|\log g(\epsilon)| = o(|\log \epsilon|)$ as $\epsilon \to 0$. One motivation for the definition at (5.2) is that in many dynamical systems, $C(\epsilon)$ exhibits power-law behaviour for small ϵ. Estimating α would be a standard statistical problem if we had a sequence of independent random variables $Y_1, Y_2, \ldots$ (constructed from the X_i's) where each random variable from the sequence was distributed as $\|U - V\|$. Nevertheless, such a construction may not be possible, even if the X_i's were independent.

Motivated by the definition at (5.2), one may consider estimating α by replacing μ in (5.1) by its empirical version, which leads to the estimator

$$\hat{\alpha}_1 = \frac{\log C_N(\epsilon_1)}{\log \epsilon_1}\,,$$

where

$$C_N(\epsilon) = \frac{2}{N(N-1)} \sum\sum_{i<j} I(\|X_i - X_j\| \leq \epsilon)\,, \tag{5.4}$$

I is the indicator function, and $\epsilon_1 > 0$. Note that if we choose ϵ_1 small, this reduces systematic bias but necessarily increases variability of $\hat{\alpha}_1$, as fewer distances are being counted. We seldom use $\hat{\alpha}_1$, however, to estimate the correlation dimension since the estimator converges only of a logarithmic rate (see for example, Theiler (1988)). To circumvent this, one may consider choosing ϵ_1 and ϵ_2 and estimating α by $\hat{\alpha}_2$, defined as

$$\hat{\alpha}_2 = \frac{\log C_N(\epsilon_2) - \log C_N(\epsilon_1)}{\log \epsilon_2 - \log \epsilon_1},$$

which improves the convergence rate. This is in essence the procedure devised by Grassberger and Procaccia (1983c), who proposed fitting a straight line to a plot of $\log C_N(\epsilon)$ and $\log \epsilon$ for a range of ϵ values. The correlation dimension, α, is estimated by the slope of the line in some region where the plot is approximately linear. Attempts to improve on this rather *ad hoc* method include using weighted least-squares, since errors in $\log C_N(\epsilon)$ are typically not identically distributed. Several authors have proved the consistency of $C_N(\epsilon)$ as an estimator of $C(\epsilon)$ under different assumptions on the sequence of $\{X_i\}$ and the underlying mechanisms that generate $\{X_i\}$; see for example Cutler (1991, 1994), Mikosch and Wang (1993) and Pesin (1993). As pointed out by Harte (1996), determining which region of the line one should use to estimate the gradient is not trivial. Often, such a region is hard to ascertain, and as a result the slope estimate becomes correspondingly arbitrary. Moreover, properties of g at (5.3) can also influence one's strategy in the choice of region.

5.3 Hill Estimator

Hill (1975) proposed a general approach to inference about the tail behaviour of a distribution, assuming it satisfies a power law. Let $Y_1, \ldots, Y_N$ be independent and identically distributed random variables from the distribution F with density f. Denote the order statistics by $Y_{(1)}, \ldots, Y_{(N)}$. We further suppose that the Y_i's satisfy

$$F(y) = P(Y \leq y) = C\, y^{\alpha} \quad \text{for} \quad y \leq \epsilon, \tag{5.5}$$

where $\alpha, C > 0$ and ϵ is known. Following standard arguments for order statistics (see David, 1980, Chapter 2), the joint density of $Y_{(1)}, \ldots, Y_{(m)}$ is

$$f_{Y_{(1)}\ldots Y_{(m)}}(y_1, \ldots, y_m) = \frac{N!}{(N-m)!}\{1 - F(y_m)\}^{N-m} f(y_1)\ldots f(y_m),$$

where $1 \leq m \leq N$ and the marginal density of $Y_{(m)}$ is

$$f_{Y_{(m)}}(y_m) = \frac{N!}{(m-1)!\,(N-m)!} F(y_m)^{m-1}\{1 - F(y_m)\}^{N-m} f(y_m),$$

and thus the conditional density of the first $m-1$ order statistics given the m'th is

$$\begin{aligned} f_{Y_{(1)}\ldots Y_{(m-1)}|Y_{(m)}}(y_1, \ldots, y_{m-1} \,|\, y_m) &= \frac{f_{Y_{(1)}\ldots Y_{(m)}}(y_1, \ldots, y_m)}{f_{Y_{(m)}}(y_m)} \\ &= (m-1)!\,\{F(y_m)\}^{1-m} f(y_1)\ldots f(y_{m-1}). \end{aligned}$$

Suppose we condition upon $Y_{(m)} \leq \epsilon$. Since F has the form (5.5), the conditional log-likelihood of α for $Y_{(1)}, \ldots, Y_{(m-1)}$, given $Y_{(m)}$, is

$$\ell_H(\alpha) = \log(m-1)! + (m-1)\log\left(\alpha Y_{(m)}^{-\alpha}\right) + (\alpha - 1)\sum_{i=1}^{m-1}\log Y_{(i)}.$$

Differentiating $\ell_H(\alpha)$ and equating to zero, the conditional maximum likelihood estimator $\hat{\alpha}_H$ of α is known as the Hill estimator, where

$$\hat{\alpha}_H = \left\{ -(m-1)\sum_{i=1}^{m-1}\log Y_{(i)} + \log Y_{(m)} \right\}^{-1}. \tag{5.6}$$

Mason (1982) showed that, for $m = m(N) \to \infty$ and $m = o(N)$, the Hill estimator is weakly consistent for α if and only if F has a regularly varying lower tail with exponent α, i.e. $\lim_{y\to 0} F(cy)/F(y) = c^\alpha$ for all $c > 0$. In the case where $m = N^\gamma$ for some $\gamma \in (0,1)$, Hall (1982) derived an optimal γ and developed asymptotic normality of $\hat{\alpha}_H$ by imposing more stringent assumptions on the tail of F. Deheuvels, Haeusler and Mason (1988) proved that whenever $m/\log\log N \to \infty$ and $m = o(N)$, the Hill estimator is strongly consistent for α.

Although the Hill estimator is backed by a number of favourable theoretical properties, they only hold when the Y_i's are independent. In practice, a given sample of interpoint distances is typically dependent. Nevertheless, Smith (1992) argued that those interpoint distances that are less than ϵ, for small ϵ, can be treated

as independent. This follows from the so-called *independent distance hypothesis* (IDH), first investigated by Theiler (1990). An argument for the IDH can be found in Smith (1992).

Mikosch and Wang (1995) proposed a refinement to the Hill estimator, and suggested using Monte Carlo method to help overcome the dependence problem. The idea is to use bootstrap methods to alleviate the impact of dependency of the data. A by-product of this procedure is an estimate of the variability of the estimates. Given a sample $\mathcal{X} = \{X_1, \ldots, X_N\}$ with common distribution F, we draw $2B$ independent resamples $\mathcal{X}_1^*, \ldots, \mathcal{X}_{2B}^*$, by sampling randomly with replacement from $\mathcal{X}$. Let $X_{1,k}^*, \ldots, X_{N,k}^*$ be the elements of the k-th resample $\mathcal{X}_k^*$. For each $1 \leq b \leq B$, we form a set of distances $\mathcal{Y}_b$ using the resamples $\mathcal{X}_{2b-1}^*$ and $\mathcal{X}_{2b}^*$ such that $\mathcal{Y}_b = \{\|X_{1,2b-1}^* - X_{1,2b}^*\|, \ldots, \|X_{N,2b-1}^* - X_{N,2b}^*\|\}$. Denote the ordered elements of $\mathcal{Y}_b$ by $Y_{(1,b)} \leq \ldots \leq Y_{(N,b)}$. For a fixed value of m, the Hill estimator $\hat{\alpha}_{H,b}$ corresponding to $\mathcal{Y}_b$ is calculated from the values $Y_{(1,b)}, \ldots, Y_{(m,b)}$ using (5.6). The final estimator of correlation dimension is given by the averages of the $\hat{\alpha}_{H,b}$'s. Mikosch and Wang (1995) proposed choosing m between $N^{1/3}$ and $N^{2/3}$, which is motivated in part by the conditions that $m = o(N)$ and $m \to \infty$. Variants of this procedure may be found in Harte (1996), which also takes boundary effects into account. Indeed, the correlation dimension estimates calculated for the earthquake data in Figure 5.2 employ a similar bootstrap procedure, with $N = 19650$ and $B = 100$. Further details are given by Harte (1996).

5.4 Takens Estimator and Binomial Estimator

We finish off this chapter by introducing two other correlation dimension estimators that exist in the literature. The Takens estimator (advocated by Takens (1985)) is similar in spirit to the Hill (1975) estimator. Both are in fact conditional maximum likelihood estimators, but with different conditioning criteria. Assume, as in the last section, that the Y_i's are independent with distribution F and (5.5) holds. Suppose that m of the Y_i's are less than ϵ, and let them be $Y_1, \ldots, Y_m$. Conditional on this, each of the m Y_i's has distribution function given by

$$F_{Y|Y\leq\epsilon}(y) = P(Y \leq y \,|\, Y \leq \epsilon) = \left(\frac{y}{\epsilon}\right)^{\alpha}, \quad 0 \leq y \leq \epsilon.$$

The corresponding log-likelihood function for α given $Y_1, \ldots, Y_m$ is

$$\ell_T(\alpha) = \sum_{i=1}^{m} \left\{ \log\left(\frac{\alpha}{\epsilon}\right) + (\alpha - 1) \log\left(\frac{Y_i}{\epsilon}\right) \right\},$$

and the conditional maximum likelihood estimator is

$$\alpha_T = \frac{m}{\sum_{i=1}^{m} \log(\epsilon / Y_i)}. \tag{5.7}$$

If we interpret the Y_i's as interpoint distances, then $\hat{\alpha}_T$ estimates the correlation dimension.

Theiler (1988) pointed out that if Y has distribution of the form $F(y) = g(y)\, y^{\alpha}$ (see also (5.3)), then $\hat{\alpha}_T$ does not necessarily converge if g is not suitably behaved. However, Theiler (1988) proved that if g satisfies

$$\lim_{\epsilon \to 0} \int_0^{\infty} \left\{ \frac{g(\epsilon e^{-w})}{g(\epsilon)} - 1 \right\} \alpha e^{-\alpha w}\, dw = 0 \tag{5.8}$$

then $\hat{\alpha}_T^{-1}$ is asymptotically unbiased for α^{-1}. In the trivial case where g is a constant, condition (5.8) holds automatically.

The binomial estimator was introduced by Smith (1992). Firstly the number of interpoint distances which are less than some pre-specified distances, say r_i, are determined. We assume that each r_j has the form $\epsilon \phi^j$, where $0 \leq j \leq m$ and $0 < \phi < 1$, and that there are N_j interpoint distances less that r_j. Let there be m interpoint distances less than ϵ, and have distribution given by (5.5). The conditional distribution of $N_1, \ldots, N_m$, given N_0, is

$$\begin{aligned} F_{N_1,\ldots,N_m|N_0}(n_1, \ldots, n_m | n_0) &= \prod_{i=1}^{m} F_{N_i|N_0,\ldots,N_{i-1}}(n_i | n_0, \ldots, n_{i-1}) \\ &= \prod_{i=1}^{m} F_{N_i|N_{i-1}}(n_i | n_{i-1}) \\ &= \prod_{i=1}^{m} \binom{n_{i-1}}{n_i} \left(\phi^{\alpha}\right)^{n_i} \left(1 - \phi^{\alpha}\right)^{n_{i-1} - n_i}, \end{aligned}$$

and the conditional log-likelihood of α is

$$\ell_B(\alpha) = \sum_{i=1}^{m} \log \binom{N_{i-1}}{N_i} + \alpha \sum_{i=1}^{m} N_i \log \phi + (N_0 - N_m) \log\left(1 - \phi^{\alpha}\right).$$

The maximum likelihood estimator of α is

$$\hat{\alpha}_B = \frac{\log\left(\sum_{i=1}^{m} N_i\right) - \log\left(\sum_{i=0}^{m-1} N_i\right)}{\log \phi}.$$

Smith (1992) termed $\hat{\alpha}_B$ the "binomial estimator", since the conditional distribution is derived using the binomial distribution. Note that this estimator does not utilise the *actual* values of interpoint distances. Finite-sample behaviour of $\hat{\alpha}_B$ was investigated in a numerical study by Smith (1992).

Chapter 6

Pole Estimation

6.1 Introduction

In this chapter, we shall consider the problem of estimating properties of poles in point-process intensity functions. We define a simple pole as an isolated point at which the intensity function is asymptotic to infinity, and a pole line as a straight or curved line-segment along which the intensity is infinite. Our work is motivated by earthquake data, which have spatial patterns of events resembling those which would arise in a point process whose intensity had a simple pole. Figure 5.1 illustrates such an example, where the epicentres in some regions are highly clustered. Our methodology is applicable more widely, and we use earthquake data only as an illustration.

We shall develop statistical methods for estimating the location and "strength" of a pole in such point process data. We define strength through the exponent, α, in a model which declares that, as x approaches a pole at the point v, the point process intensity at x satisfies $\Lambda(x) \sim \text{const.} \|x - v\|^{-\alpha}$. A similar definition applies in the case of pole lines.

Sometimes, a pole line in images such as that in the lower left-hand corner of Figure 5.1 is formed by a pole migrating with time. In other words, if the data were recorded as a time series, the line would appear as a sequence of isolated points. In such cases, it may be more appropriate and informative to analyse the data as a time series, and chart the movement of the pole with time. Our methods allow us to perform this type of analysis.

As a first approximation to data such as those in Figure 5.1, and on grounds of

simplicity and plausibility, we suggest Poisson process models for epicentres. These have been implicitly adopted by authors who have analysed earthquake data in the fractal context. See for example Harte (1996) and especially Section 6.4, where we shall show how the strength of a pole may be related to the correlation dimension of a point process. Even when Poisson models are not entirely correct, these approaches nevertheless motivate methods and estimators that are generally valid, at least in terms of statistical consistency. We shall explore basic theoretical properties of our estimators in terms of bias, variance and convergence rate. Note that, in deriving estimators of correlation dimension in Chapter 5, the same philosophy was applied by assuming independent interpoint distances, and statistical consistency may be achieved for those methods when the point processes are at least weakly dependent. Methods of statistical inference for general point process models may be found in monographs by Ripley (1981) and Cressie (1993).

In the treatment of pole estimation in this chapter, we shall confine ourselves mainly to two-dimensional data, and generalisations to higher dimensions will be only briefly mentioned. Although these extensions are relatively straightforward, they seem to be unmotivated by applications. Since we are concentrating on point process data, we shall treat explicitly only the case of estimating poles in intensity functions. In the context of density functions, particularly for independent data, the methods are virtually identical to those treated here.

Our methods are semiparametric in nature, requiring only "asymptotic" models for the intensity. Although we shall use structural models to suggest techniques, we show that the techniques lead to accurate estimators under general assumptions. We shall see in Section 6.7 that existing methods for estimating pole strength, for example using fractal properties, are restrictive in terms of the range of strengths that they allow. They are statistically consistent only in that half of the range which corresponds to relatively strong poles, and cope poorly with poles whose strengths lie in the middle range. Our analysis of real data in Section 6.8 shows that in practice, pole strengths often lie in the middle range.

This chapter is organised as follows. Section 6.2 reviews basic properties of Poisson point process. Sections 6.3 and 6.4 introduce our methodology for estimating properties of poles using maximum likelihood techniques and nonparametric methods, respectively. Pole-line estimation is discussed in Section 6.5, and modified techniques for dealing with observational errors are detailed in Section 6.6. Theoretical

properties and numerical studies are summarised in Sections 6.7 and 6.8.

6.2 Poisson Process Properties

Let $X_1, \ldots, X_N$ denote the observations of a point process $\mathcal{P}$ in a given region $\mathcal{R} \subseteq \mathbb{R}^d$. We call $\mathcal{P}$ a Poisson process, with intensity function $\Lambda \geq 0$, if (i) the number of points X_i in any Borel subset $\mathcal{S}$ of $\mathbb{R}^d$ is Poisson-distributed with mean $\int_{\mathcal{S}} \Lambda$, and (ii) the numbers of points in any finite number of disjoint Borel subsets of $\mathbb{R}^d$ are independent random variables. If Λ is constant almost everywhere, then $\mathcal{P}$ is called a *homogeneous* or *uniform* Poisson process.

Poisson processes are important building blocks of more complicated models. An example is the *Neyman-Scott process*, where the so-called parent points come from a Poisson process, and daughter points are distributed independently around each parent point. See Ripley (1981, p. 164ff) for discussions of such models.

Following directly from condition (i), the mean and variance of the number of points in $\mathcal{S}$ equal $\int_{\mathcal{S}} \Lambda$, and the probability of having no points in $\mathcal{S}$ is $\exp(-\int_{\mathcal{S}} \Lambda)$. A useful property of a Poisson process is that, conditional on there being N points in $\mathcal{S}$, these are independent and identically distributed, with density $\lambda(x) = \Lambda(x)/\int_{\mathcal{S}} \Lambda$ for $x \in \mathcal{S}$. In particular, if $\mathcal{P}$ is homogeneous then the N observations are independent and uniformly distributed in $\mathcal{S}$.

We shall prove this property for the case $d = 1$ (see Cox and Isham, 1980, Chapter 3). Suppose, without loss of generality, that there are M points lying in the interval $\mathcal{I}_0 = (0, \xi]$ from a Poisson process with intensity Λ. Let $0 = \xi_0 < \xi_1 < \ldots < \xi_M \leq \xi$ and $\delta_1, \ldots, \delta_M > 0$. Denote by $N(\mathcal{I})$ the number of points lying in the interval $\mathcal{I}$, and by $\mathcal{J}_i$ the interval $[\xi_i, \xi_i + \delta_i)$. Then,

$$\begin{aligned}
&\lim_{\delta_1,\ldots,\delta_M \to 0} (\delta_1 \ldots \delta_M)^{-1} P\{N(\mathcal{I}_0) = M,\ N(\mathcal{J}_1) = 1, \ldots,\ N(\mathcal{J}_M) = 1\} \\
&= \left[\prod_{i=1}^{M} \Lambda(\xi_i) \exp\left\{-\int_{\xi_{i-1}}^{\xi_i} \Lambda(x)\,dx\right\}\right] \exp\left\{-\int_{\xi_M}^{\xi} \Lambda(x)\,dx\right\} \\
&= \left\{\int_{\mathcal{I}_0} \Lambda(x)\,dx\right\} \prod_{i=1}^{M} \Lambda(\xi_i)\,,
\end{aligned}$$

whence it follows that

$$\lim_{\delta_1,\ldots,\delta_M \to 0} (\delta_1 \ldots \delta_M)^{-1} P\{N(\mathcal{J}_1) = 1, \ldots,\ N(\mathcal{J}_M) = 1 \mid N(\mathcal{I}_0) = M\}$$

$$= \left[P\{N(\mathcal{I}_0) = M\}\right]^{-1} \left\{ \int_{\mathcal{I}_0} \Lambda(x)\, dx \right\} \prod_{i=1}^{M} \Lambda(\xi_i)$$

$$= M! \prod_{i=1}^{M} \left\{ \Lambda(\xi_i) \Big/ \int_{\mathcal{I}_0} \Lambda(x)\, dx \right\},$$

which is the joint density for the order statistics of M independent random variables each with density function $\Lambda / \int_{\mathcal{I}_0} \Lambda$. The monograph by Kingman (1993) gives detailed discussions of Poisson processes.

6.3 Maximum Likelihood Estimation

In this section we investigate maximum likelihood techniques for estimating the parameters of our simple model. We assume that the point process data X_i are observed within a subset $\mathcal{R}$ of the plane, and write Λ for the intensity function of the point process defined on $\mathcal{R}$.

Suppose that Λ has a simple pole at $v \in \mathcal{R}$, and that the pole is approached "at rate α", in the sense that, in some neighbourhood of v,

$$\Lambda(x) = C \, \|x - v\|^{-\alpha}, \tag{6.1}$$

where $\alpha, C > 0$ and we take $\| \cdot \|$ to be the Euclidean metric. In order for the expected number of points in each bounded, non-degenerate region to be finite, it is not difficult to see that we have to impose the condition $\alpha < 2$. Likewise, in the d-dimensional case, we require $\alpha < d$. Of course, if the point process were Poisson then the actual number of data in a region would be infinite, with probability 1, if the expected number there were infinite.

If the point process were Poisson, we might consider estimating v and α by maximum likelihood, as follows. Suppose that just N of the X_i's lie in a given region $\mathcal{R} \subseteq \mathbb{R}^2$ containing the unknown v, and we denote them by $X_1, \ldots, X_N$. Conditional on these N values, the likelihood of the X_i's is

$$L = \prod_{i=1}^{N} \frac{\|X_i - v\|^{-\alpha}}{\int_{\mathcal{R}} \|x - v\|^{-\alpha}\, dx}. \tag{6.2}$$

Define the log-likelihood to be $\ell(v, \alpha) = -\log L$. Clearly, $L = +\infty$ if $v = X_i$ for some $1 \le i \le N$. Therefore, estimating v by maximum likelihood method is not feasible.

However, for any v not equal to one of the X_i's it may be shown that a maximum likelihood estimator for α exists. This follows from properties of the partial derivatives of $\ell(v,\alpha)$ with respect to α, as follows. The first partial derivative is given by

$$\Big(\frac{\partial}{\partial\alpha}\Big)\,\ell(v,\alpha) = \sum_{i=1}^{N} \log\|X_i - v\| - N\,\frac{\int_{\mathcal{R}} \|x-v\|^{-\alpha}\log\|x-v\|\,dx}{\int_{\mathcal{R}}\|x-v\|^{-\alpha}\,dx}\,. \tag{6.3}$$

As $\alpha \to \infty$, the second term on the right hand side of (6.3) diverges to $+\infty$, and so too does $(\partial/\partial\alpha)\,\ell(v,\alpha)$. On the other hand, as $\alpha \to -\infty$, $(\partial/\partial\alpha)\,\ell(v,\alpha)$ converges to a finite, negative number, since the right hand side of (6.3) converges to $\sum_{i=1}^{N}(\log\|X_i - v\| - \log\|x_0 - v\|)$ for some x_0 on the boundary of $\mathcal{R}$ such that the distance between x_0 and v is maximised. Also, from the Cauchy-Schwarz inequality, the second derivative of $\ell(v,\alpha)$, i.e.

$$\begin{aligned}\Big(\frac{\partial^2}{\partial\alpha^2}\Big)\,\ell(v,\alpha) = -N\Big(\int_{\mathcal{R}}\|x-v\|^{-\alpha}\,dx\Big)^{-2}\Big[\int_{\mathcal{R}}\big(\|x-v\|^{-\alpha}\log\|x-v\|\big)^2\,dx \\ -\Big(\int_{\mathcal{R}}\|x-v\|^{-\alpha}\,dx\Big)\Big\{\int_{\mathcal{R}}\|x-v\|^{-\alpha}\big(\log\|x-v\|\big)^2\,dx\Big\}\Big]\,,\end{aligned}$$

is greater than 0 for each α. Therefore, given an estimator $\hat{v}$ of v, not equal to one of $X_1,\ldots,X_N$, the equation $(\partial/\partial\alpha)\,\ell(\hat{v},\alpha) = 0$ has a unique solution $\hat{\alpha}$, which we may take as an estimator of α.

In practice, to avoid numerical problems that arise from $\hat{v}$ being too close to one or more of the X_i's, when defining $\hat{\alpha}$ it may be advisable to remove from the data those X_i's that lie in the very near vicinity of $\hat{v}$. Furthermore, since the model $\Lambda(x) = C\,\|x-v\|^{-\alpha}$ is usually only correct in an approximate sense (as x converges to v), we may reduce the effect of bias due to model selection error by replacing $\mathcal{R}$ by a small region in the vicinity of $\hat{v}$. Therefore, one might define $\hat{\alpha}$ as the minimiser of

$$\ell(\alpha) = \alpha\sum{}' \log\|X_i - \hat{v}\| + M(\mathcal{S}_2\backslash\mathcal{S}_1)\,\log\Big(\int_{\mathcal{S}_2\backslash\mathcal{S}_1}\|x-\hat{v}\|^{-\alpha}\,dx\Big)\,, \tag{6.4}$$

where $\mathcal{S}_1 \subseteq \mathcal{S}_2$ are concentric discs centred on $\hat{v}$, $\sum'$ denotes summation over those points X_i that lie in $\mathcal{S}_2\backslash\mathcal{S}_1$, and $M(\mathcal{S}_2\backslash\mathcal{S}_1)$ equals the number of such points. As before, $\hat{\alpha}$ is uniquely defined by this prescription. The extension to d-dimensional data is entirely analogous (e.g. by changing $\mathcal{S}_1, \mathcal{S}_2$ to d-dimensional spheres). The radius r_2 of $\mathcal{S}_2$ plays the role of a smoothing parameter, in the sense that choosing

larger r_2 reduces variance but (if the model is only approximately correct) increases bias. Theoretical properties of $\hat{\alpha}$ will be given in Section 6.7, where we shall derive the rate of convergence of $\hat{\alpha}$ to α. Note that in our numerical studies, we have not explicitly removed points that are too close to the estimated pole $\hat{v}$. There are two reasons. Firstly, the sample size we used is typically small, and thus computational difficulties arising from very small distances between X_i's and $\hat{v}$ are minor. Second, using an annulus requires the choice of an inner radius r_1 for $\mathcal{S}_1$, which has to be selected jointly with r_2. Our experience is that the estimator becomes relatively unstable for a poor choice of r_1.

6.4 Nonparametric Estimation

6.4.1 Pole Location

We consider estimating v by maximising the number of points within a small region. Let $\mathcal{S}(w, r)$ denote the closed disc of radius r centred at $w \in \mathbb{R}^2$. We may define $\hat{v}$ to be a value of w which maximises the number of points contained in $\mathcal{S}(w, r)$ for a given value of r. Alternatively, we may define $\hat{v}$ as the value of w which minimises the area of $\mathcal{S}(w, r)$ subject to the disc containing at least a given number, say m, points. If the points X_i are distributed in the continuum, then the former $\hat{v}$ is not uniquely defined *with probability 1*, while the latter is unique, with the same probability. Figure 6.1 demonstrates the former case graphically. Therefore, in our theoretical and numerical studies, we employ the latter definition of $\hat{v}$ for estimating pole locations. Extension to higher dimensions is obvious.

6.4.2 Pole Strength

There are several approaches available for estimating α. One is based on the parametric prescription at (6.4), by constructing the estimator $\hat{v}$ given in Section 6.4.1, and substituting it for the definition of v on the right-hand side of (6.4), and defining $\hat{\alpha} = \operatorname{argmin} \ell(\alpha)$. As we shall see in Theorem 6.3, this estimator achieves consistency for a wide range of values of α, in particular for $0 < \alpha < 2$. It is not necessary for the model generating the data to be precisely that used to produce the maximum likelihood equation (6.4).

Alternatively, one may circumvent estimating v and base inference instead on

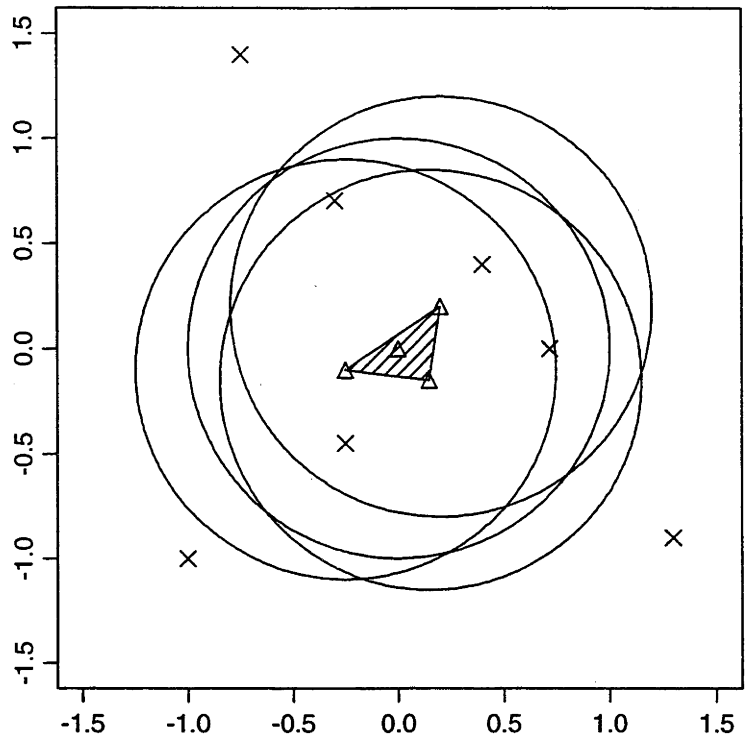


Figure 6.1: An example demonstrating non-uniqueness of one type of pole location estimator. Here we estimate v as the point w that maximises the number of points contained in $\mathcal{S}(w, r)$ for a given value of r. The crosses and triangles denote data points and centres of the four circles respectively. Each circle has fixed radius $r = 1$, and includes a maximum number of points. Indeed, any point lying in the shaded region is a possible estimate for the pole location.

interpoint distances $\|X_i - X_j\|$, for example by using methods of Hill (1975) or Grassberger and Procaccia (1983) discussed in Chapter 5. These are in fact motivated by inferential problems for independent and identically distributed scalar random variables whose density has a pole at the origin, although (depending on the value of α) they are readily adapted to problems of inference for the non-independent variables $\|X_i - X_j\|$. Ripley (1981, Chapter 8) showed how to estimate the intensity of a homogeneous Poisson process using interpoint distances.

If the intensity Λ of the Poisson process generating the points X_i has a pole of order α at v (that is, if $\Lambda(x) \sim C \|x - v\|^{-\alpha}$ as $x \to v$), then first-order properties of the distribution of $\|X_i - X_j\|$ near the origin depend on α if and only if $\alpha > 1$.

(In the d-dimensional case, the condition changes to $\alpha > \frac{1}{2}d$). Indeed, if $\alpha < 1$ then $p(u) \equiv P(\|X_i - X_j\| \le u) \sim Cu^2$ as $u \to 0$, where C denotes a generic positive constant. (Here we assume that X_i and X_j are arbitrary distinct points of the Poisson process in a neighbourhood of v.) For $1 < \alpha < 2$, $p(u) \sim Cu^{2(2-\alpha)}$, and for $\alpha = 1$, $p(u) \sim Cu^2|\log u|$ as $u \to 0$. These facts follow from (6.42) and (6.44) in the proof of Theorem 6.2.

Thus, it is only when $1 < \alpha < 2$ that we can expect to estimate α consistently by the Hill and Grassberger–Procaccia approaches. Those techniques break down completely when $\alpha \le 1$. By way of comparison, the approach based on (6.4) produces consistency whenever $0 < \alpha < 2$, and under quite general models. Nevertheless, for $\alpha \in (1, 2)$ the technique based on $\|X_i - X_j\|$ has reasonable convergence properties, and variants of it have been considered by Mikosch and Wang (1995), Harte (1996) and Vere-Jones (1996) in the context of inference about fractal properties of a point process.

Specifically, define $\gamma = \{2(2-\alpha)\}^{-1}$, $N(u) = \sum\sum_{i<j} I(\|X_i - X_j\| \le u)$,

$$\bar{u} = u^{-1}(1-\theta)^{-1} \int_{\theta u}^{u} \log t \, dt\,, \quad I(u) = \int_{\theta u}^{u} (\log t - \bar{u})^2 \, dt\,,$$

$$\tilde{\gamma} = \tilde{\gamma}(u) = N(u)^{-1} \sum\sum_{i<j} \left(\log u - \log \|X_i - X_j\|\right) I\left(\|X_i - X_j\| \le u\right),$$

$$\check{\gamma} = \check{\gamma}(u) = I(u)^{-1} \int_{\theta u}^{u} (\log t - \bar{u}) \log N(t) \, dt\,.$$

Both $\tilde{\gamma}$ and $\check{\gamma}$ consistently estimate γ, and so $\tilde{\alpha} = 2 - (2\tilde{\gamma})^{-1}$ and $\check{\alpha} = 2 - (2\check{\gamma})^{-1}$ are consistent for α. See Theorem 6.2. Note that $\tilde{\gamma}$ is essentially the Takens estimator derived in Section 5.4. The quantity $u > 0$ is a smoothing parameter, and $0 < \theta < 1$. We would generally take θ close to zero, although since $N(u) = 0$ in a neighbourhood of the origin we cannot allow $\theta = 0$. Similar remarks can be made in d-dimensional cases, where $\tilde{\alpha} = d - (2\tilde{\gamma})^{-1}$ and $\check{\alpha} = d - (2\check{\gamma})^{-1}$ are consistent for α.

In Section 6.7 we shall show that the estimators $\hat{\alpha}, \tilde{\alpha}, \check{\alpha}$ are generally consistent for α, and have convergence rates that are polynomially fast as functions of the "average" intensity of the point process.

In the context of Chapter 5, γ^{-1} is termed the correlation dimension, D, of the Poisson point process, and is simply related to the strength of a pole by the following formula:

$$\alpha = 2 - \frac{D}{2}\,. \tag{6.5}$$

We argue that there is much to be gained by considering this estimation problem in non-fractal terms, and our approach provides a relatively simple context in which to interpret features of the point pattern. In particular, for a relatively strong pole, the correlation dimension of a point process will be close to 0.

6.5 Estimation of Pole Line

Let $\mathcal{C} \subseteq \mathbb{R}^2$ be a curve of finite length, with the property that $\Lambda(x) = \infty$ along $\mathcal{C}$. We call $\mathcal{C}$ a pole line. In data on earthquake catalogues such as those in Section 6.8, poles have lengths of a kilometre or so, and represent subterranean geological faults. To the resolution of the data, they usually appear as points. But in some epicentre catalogues they are visible as line segments. We shall show how some of our earlier methods may be modified to estimate pole lines.

An approximate model for the intensity function in a neighbourhood of a pole line, $\mathcal{C}$, may be defined in an analogous manner as for the intensity function with a pole at (6.1), and is given by

$$\Lambda(x) = C \sup_{v \in \mathcal{C}} \|x - v\|^{-\alpha}. \tag{6.6}$$

The expected number of points in the neighbourhood of $\mathcal{C}$ is infinite unless $0 < \alpha < 1$.

As with the model at (6.1), the simple model at (6.6) serves only to motivate methods which produce consistent estimators in more general contexts. It does, however, imply that the intensity diverges at a constant, polynomial rate along the pole line. Introducing a variable rate seems awkward, without producing unattractively complex methods. Unless the application requires such complexities, those methods do not seem justified in practice, especially if the pole lines are short.

We may estimate $\mathcal{C}$ and α using modified versions of the methods suggested in Sections 6.3 and 6.4. Let $\mathcal{S}(w, r)$ be the closed disc centred at w, and $\mathcal{W}$ be the set of all points w in the vicinity of $\mathcal{C}$ such that (a) $\mathcal{S}(w, r)$ has 3 points X_i on its boundary, (b) $\mathcal{S}(w, r)$ contains at least m points of the Poisson process, and (c) $r \leq r_0$, where $m \geq 1$ and $r_0 > 0$ are given constants. Condition (a) implies that for a given centre w, r is as small as possible subject to $\mathcal{S}(w, r)$ satisfying (b). In practice, the "vicinity" of $\mathcal{C}$ is easy to determine visually. With each $w \in \mathcal{W}$ we associate the unique value r_w representing the radius of the closed disc $\mathcal{S}(w, r_w)$. Using the set of pairs $\{(w, r_w) : w \in \mathcal{W}\}$ we may fit a curve $\widehat{\mathcal{C}}$ to $\mathcal{C}$ either parametrically or

nonparametrically.

For example, if our model for $\mathcal{C} = \mathcal{C}(a, b)$ is a straight-line segment with equation $y = ax + b$, let $D(w; a, b)$ be the perpendicular distance from $w \in \mathcal{W}$ to $\mathcal{C}(a, b)$, and choose $(\hat{a}, \hat{b})$ to minimise

$$\sum_{w \in \mathcal{W}} D(w; a, b)\, A(r_w)\,,$$

where $A(\cdot)$ is a non-increasing function. Fitting non-linear parametric models for pole lines, and fitting pole lines by nonparametric local linear techniques (such as those described in Chapter 1), involve similar arguments. The length of a pole line may be determined either visually or, more objectively, directly from the data. For example, we may decide that the pole line extends only over a range such that each part of it is no more than a given distance from at least one point X_i.

Having determined $\widehat{\mathcal{C}}$ we may estimate α by modifying the estimator defined at (6.4), as follows. Given $u > 0$ and $\theta \in (0, 1)$, let $\mathcal{T}_1$ and $\mathcal{T}_2$ denote the sets of all points in $\mathbb{R}^2$ that lie no further than θu and u from $\widehat{\mathcal{C}}$, respectively. Put $\widehat{\Lambda}(x|\alpha) = \sup_{v \in \widehat{\mathcal{C}}} \|x - v\|^{-\alpha}$, and let $\hat{\alpha}$ be the unique minimiser of

$$\sum\nolimits'' \left[-\log \widehat{\Lambda}(X_i|\alpha) + \log \left\{ \int_{\mathcal{T}_2 \setminus \mathcal{T}_1} \widehat{\Lambda}(x|\alpha)\, dx \right\} \right], \tag{6.7}$$

where $\sum''$ denotes summation over all points $X_i \in \mathcal{T}_2 \setminus \mathcal{T}_1$. It may be shown that, under models approximate to that at (6.6), $\hat{\alpha}$ is consistent for α in the range $0 < \alpha < 1$. The estimators $\tilde{\alpha}$ and $\check{\alpha}$ suggested in Section 6.4 are typically not appropriate in the case of pole lines, since first-order properties of the distribution of $\|X_i - X_j\|$ near the origin do not depend, to first order, on α. We do not investigate the problem of estimating pole lines in our simulation studies, since minimising (6.7) by numerical means is extremely computationally intensive.

6.6 Sources of Error

Although the Kanto earthquake data to which we shall apply our methods are of relatively high quality (for example, compared to New Zealand earthquake catalogue), they still suffer from several possible sources of error. These include stochastic epicentre location error, which has more prominent effect on those epicentres close to a pole. It may be modelled by the addition of independent random vectors (with zero means) to hypothetical "true" earthquake centre measurements. For the sake

of simplicity we assume that these vectors are identically distributed, even though the errors may not be uniform over any individual analysed region. A model for this will be developed in the next paragraph. There is also systematic error in epicentre approximation, for example errors that arise from using poorly placed ground stations. In our numerical work we avoid this problem by confining attention to data from areas which are well served by recording stations, and so do not consider offshore or coastal regions where detectability is comparatively low. Bayside regions do not suffer from the same problems, provided seismic stations are available on the opposite side of the bay. Additionally, the epicentres of our data were rounded to three decimal places of degrees of latitude and longitude. We assume that our analysed region is flat and ignore curvature of the earth. This should not greatly affect our results since the regions of interest are typically small.

To take stochastic errors into account, we modified the likelihood L at (6.2) by convolving it with a bivariate error distribution. We took the latter to be spherically symmetric and Normally distributed, obtaining the following analogue of L:

$$L_1(X_1, \ldots, X_N | \, v, \alpha, \sigma) = \prod_{i=1}^{N} \frac{\Lambda_1(X_i | \, v, \alpha, \sigma)}{\int_{\mathcal{S}} \Lambda_1(x | \, v, \alpha, \sigma) \, dx} , \tag{6.8}$$

where $\Lambda_1(x|v, \alpha, \sigma) = \int \|x + \sigma y - v\|^{-\alpha} \, \phi(y) \, dy$, ϕ is the standard bivariate Normal $N(0, I)$ density, $\mathcal{S}$ is a suitable region where the observations lie, and $\sigma > 0$. The analogue, ℓ_1, of the negative log-likelihood, ℓ, defined at (6.4), may be defined similarly. Of course, one may develop more sophisticated models, say to accommodate the covariance structure of the two components of noise, but it necessarily complicates the estimation procedure. Moreover, as we shall see in our numerical studies, there is very little information in the data for estimating σ using the model Λ_1, and maximising L_1 (or minimising ℓ_1) over both α and σ (with $\hat{v}$ replacing v) is not a practical option. Nevertheless, by varying σ one can assess the way in which estimates alter for different amounts of noise.

6.7 Large-Sample Theory

We shall establish consistency and rates of convergence of our estimators of pole location and pole strengths discussed in earlier sections. Suppose we observe points of a Poisson process with intensity $\Lambda = \nu\lambda$ in a plane, where λ is a fixed intensity

function and ν is a scalar whose value we shall allow to diverge. Let there be just N points, $X_1, \ldots, X_N$ say, in a given region $\mathcal{R} \subseteq \mathbb{R}^2$. Conditional on this event, $X_1, \ldots, X_N$ are the values of N independent and identically distributed random variables with density $f(x) = c^{-1}\lambda(x)$, for $x \in \mathcal{R}$, where $c = \int_{\mathcal{R}} \lambda$ (see Section 6.2). Allowing ν to diverge ensures that N does too. Assume for definiteness that $\mathcal{R}$ is a disc, of which v is an interior point.

First we consider properties of $\hat{v}$, defined in Section 6.3. We ask that $\lambda(x)$ behave like the function $\|x - v\|^{-\alpha}$, in the sense that

$$C_1 \leq \inf_{x \in \mathcal{R}} \lambda(x) \|x - v\|^{\alpha} \leq \sup_{x \in \mathcal{R}} \lambda(x) \|x - v\|^{\alpha} \leq C_2 \tag{6.9}$$

for constants $0 < C_1 \leq C_2 < \infty$. Given $C_3 > 0$, let $m = m(\nu)$ denote any sequence of integers such that

$$C_3 \log \nu \leq m = o(\nu)\,, \tag{6.10}$$

and define $\hat{v}$ to be the centre of that disc within $\mathcal{R}$ whose area is smallest possible subject to containing at least m points of the Poisson process.

Theorem 6.1 *Assume condition (6.9), and that $0 < \alpha < 2$. Then, provided C_3 is chosen sufficiently large and satisfies (6.10) , there exists a constant $C_4 > 0$ such that*

$$P\{\|\hat{v} - v\| \leq C_4(m/\nu)^{1/(2-\alpha)}\} \to 1\,.$$

Remark 6.1. *Rate of convergence of $\hat{v}$ to v.* In view of Theorem 6.1, the rate of convergence of $\hat{v} - v$ to zero is at least $O_p\{(m/\nu)^{1/(2-\alpha)}\}$ as ν increases. Since we may select m as small as a constant multiple of $\log \nu$ then the convergence rate given by Theorem 6.1 is as fast as $(\nu^{-1} \log \nu)^{1/(2-\alpha)}$. Note that this rate is always faster than $\nu^{-1/2}$, and is better than ν^{-C} (for any given $C > 0$) whenever α is sufficiently close to 2.

Proof of Theorem 6.1. Let $\mathcal{R}'$ denote any closed disc contained in the interior of $\mathcal{R}$ and containing v as an interior point. We shall derive the result in the case where $\hat{v}$ is defined by taking the supremum over discs in $\mathcal{R}'$ instead of over discs in $\mathcal{R}$. A relatively simple subsidiary argument allows us to remove this restriction. We need two lemmas in our proof. The first is a version of Bernstein's inequality (see Pollard, 1984, p. 191ff) and the second gives exponential bounds on Poisson

tail probabilities. To be consistent, the constants $C_1, \ldots, C_4$ are those as stated in the theorem.

Lemma 6.1. *Let $Z_1, \ldots, Z_N$ be independent random variables. Suppose $P\{|Z_i - E(Z_i)| \leq m\} = 1$ for each i where $m < \infty$. Then, for $s > 0$,*

$$P\left[\left|\sum_{i=1}^{N}\{Z_i - E(Z_i)\}\right| \geq Ns\right] \leq 2\exp\left\{-\frac{N^2s^2}{2\sum_{i=1}^{N}\operatorname{var}(Z_i) + \frac{2}{3}Nms}\right\}. \quad (6.11)$$

A proof of Lemma 6.1 may be found in Pollard (1984, p. 191ff). Note that if the random variables Z_i have constant variance, say $\operatorname{var}(Z_i) = \sigma^2$, the upper bound reduces to

$$2\exp\left(-\frac{Ns^2}{2\sigma^2 + \frac{2}{3}ms}\right).$$

Lemma 6.2. *Let Z be a Poisson random variable with mean λ. For each $s > 0$ in the "+" case, and each $0 < s \leq 1$ in the "−" case,*

$$P\left\{\pm\frac{(Z-\lambda)}{\sqrt{\lambda}} \geq s\right\} \leq \exp\left\{-\frac{s^2}{2}\psi\left(\frac{\pm s}{\sqrt{\lambda}}\right)\right\}, \quad (6.12)$$

where $\psi(s) = 2h(1+s)/s^2$ and $h(s) = s(\log s - 1) + 1$.

Note in particular that $\psi(s) = 1 - \frac{1}{3}s + o(s)$ for $s \to 0$. The monograph by Shorack and Wellner (1986, Chapter 11) gives detailed results on bounds for Poisson random variables.

Conditional on N, the variables $X_1, \ldots, X_N$ are independent and identically distributed with density $f = c^{-1}\lambda$, and the number $M(\mathcal{S})$ of them lying within $\mathcal{S}$ is binomial $\mathrm{Bi}\{N, p(\mathcal{S})\}$, where $p(\mathcal{S}) = \int_{\mathcal{S}} f$. Putting $m = 2$ and $s = t\{N^{-1}p(\mathcal{S})\}^{1/2}$ in (6.11), we see that given any $\epsilon \in (0,1)$, there exists a constant $C > 0$, depending only on ϵ, such that for all $t > 0$ and all sets $\mathcal{S}$ satisfying $p(\mathcal{S}) \leq 1 - \epsilon$,

$$P\left[|M(\mathcal{S}) - N\,p(\mathcal{S})| > t\{N\,p(\mathcal{S})\}^{1/2}\middle|N\right] \leq 2\exp\left(-Ct\min\left[t, \{Np(\mathcal{S})\}^{1/2}\right]\right).$$

Note too that on replacing s by $B(c^{-1}\log\nu)^{1/2}$ in (6.12), we have

$$\begin{aligned} P\left\{\left|\frac{N-c\nu}{\sqrt{c\nu}}\right| \geq B(c^{-1}\log\nu)^{1/2}\right\} &\leq 2\exp\left[-\frac{B^2c^{-1}\log\nu}{2}\psi\left\{\frac{\pm B(c^{-1}\log\nu)^{1/2}}{\sqrt{c\nu}}\right\}\right] \\ &= 2\exp\left(\log\nu^{-B^2/2c}\left[1 + O\{(\nu^{-1}\log\nu)^{1/2}\}\right]\right) \\ &= O(\nu^{-A}), \end{aligned}$$

and hence for each $A > 0$ there exists $B = B(A) > 0$ such that

$$P\{|N - c\nu| > B\,(\nu \log \nu)^{1/2}\} = O(\nu^{-A})\,, \tag{6.13}$$

where $c = \int_{\mathcal{R}} \lambda$. Therefore, if $\epsilon \in (0,1)$ and $\mathcal{A} = \mathcal{A}(\nu)$ denotes any class of discs $\mathcal{S}$ containing no more than ν^A elements, for some fixed $A > 1$, and such that

$$\nu^{-1} \log \nu \le p(\mathcal{S}) \le 1 - \epsilon \tag{6.14}$$

for each $\mathcal{S} \in \mathcal{A}$; then, for $B = B(A) > 1$ sufficiently large,

$$\begin{aligned}
&P\big[|M(\mathcal{S}) - N\,p(\mathcal{S})| > B\,\{\nu\,p(\mathcal{S})\,\log\nu\}^{1/2} \text{ for some } \mathcal{S} \in \mathcal{A}\big] \\
&\le \nu^A \sum_{k=0}^{\infty} P\big[|M(\mathcal{S}) - N\,p(\mathcal{S})| > B\,\{\nu\,p(\mathcal{S}) \log\nu\}^{1/2} \,\big|\, N = k\big]\, P(N = k) \\
&\le \nu^A \Big(P\{|N - c\nu| > B\,(\nu \log \nu)^{1/2}\} \\
&\qquad + \sum_k{}' P\big[|M(\mathcal{S}) - k\,p(\mathcal{S})| > B\big(k^{-1}\nu \log\nu\big)^{1/2}\{k\,p(\mathcal{S})\}^{1/2}\big]\, P(N = k)\Big) \\
&\le \nu^A \Big\{ P\{|N - c\nu| > B\,(\nu \log \nu)^{1/2}\} \\
&\qquad + 2 \sum_k{}' \exp\Big(- C_5\,\big(k^{-1}\nu \log\nu\big)^{1/2} \min\big[B\,\big(k^{-1}\nu\log\nu\big)^{1/2}, \{k\,p(\mathcal{S})\}^{1/2}\big]\Big) \\
&\qquad\qquad \times P(N = k)\Big\} \\
&\le 2\,\nu^A \big\{ \exp(-BC_6 \log\nu - C_7) + o\big(\nu^{-A}\big)\big\} \to 0
\end{aligned} \tag{6.15}$$

as $\nu \to \infty$, where C_5, C_6, C_7 are positive constants which depend only on B and ϵ, $C_6 > AB^{-1}$, and $\sum_k{}'$ denotes summation over

$$\max\{\lfloor c\nu - B(\nu \log \nu)^{1/2} \rfloor,\, 0\} \le k \le \lceil c\nu + B(\nu \log \nu)^{1/2} \rceil\,.$$

Let r_0 denote the radius of $\mathcal{R}$. Given $A_1 > 0$, let $\mathcal{L}$ be a square lattice of points in $\mathbb{R}^2$ with edge width ν^{-A_1}; let $\mathcal{M}$ be the set of members of the sequence $r = i\nu^{-A_1}$, $i \ge 1$, such that $r \le r_0$; let $\mathcal{A} = \mathcal{A}(\nu)$ be the set of all discs that are centred at a point of $\mathcal{L}$, have radius in $\mathcal{M}$, satisfy (6.14) and are contained in $\mathcal{R}$; and let $\mathcal{B} = \mathcal{B}(\nu)$ be the set of all discs that have radius r satisfying $\nu^{-A_1/2} \le r \le r_0$, satisfy (6.14) and are contained in $\mathcal{R}'$. Noting that the boundary of $\mathcal{R}'$ is bounded away from that of $\mathcal{R}$ we see that we may choose ϵ in (6.14) so small that the second inequality there is satisfied for all $\mathcal{S} \subseteq \mathcal{R}'$. Now, $\mathcal{A}$ has no more than ν^A elements, for some

$A = A(A_1) > 0$, and so (6.15) applies to $\mathcal{A}$. Given $A_2 > 0$ we may choose A_1 so large that, for all sufficiently large ν, for each $\mathcal{S} \in \mathcal{B}$ there exist $\mathcal{S}_1, \mathcal{S}_2 \in \mathcal{A}$ with the property that $\mathcal{S}_1 \subseteq \mathcal{S} \subseteq \mathcal{S}_2$ and $p(\mathcal{S}_2) \leq p(\mathcal{S}_1)\,(1 + \nu^{-A_2})$. Therefore, noting (6.13) we see that since (6.15) applies to $\mathcal{A}$ it must also apply to $\mathcal{B}$. (Note that $M(\mathcal{S}_1) \leq M(\mathcal{S}) \leq M(\mathcal{S}_2)$.)

Given $C_7 > 0$, define

$$\mathcal{D}_1(C_7) = \{\text{discs } \mathcal{S} \subseteq \mathcal{R}' \text{ such that } p(\mathcal{S}) \geq C_7 \nu^{-1} \log \nu\}\,.$$

From (6.13), and the version of (6.15) with $\mathcal{A}$ replaced by $\mathcal{B}$, we see that

$$P\Big[|\mathcal{M}(\mathcal{S}) - c\nu p(\mathcal{S})| \leq B\{p(\mathcal{S})\,\nu \log \nu\}^{1/2}\{1 + p(\mathcal{S})^{1/2}\} \\ \text{for all discs } \mathcal{S} \in \mathcal{D}_1(C_7)\Big] \to 1\,.$$

Choosing $C_7 \geq (4c^{-1}B)^2$, we may show that $16B^2\nu^{-1} \log \nu \leq c^2 p(\mathcal{S})$ for $\mathcal{S} \in \mathcal{D}_1(C_7)$, which implies $B\{p(\mathcal{S})\nu \log \nu\}^{1/2}\{1 + p(\mathcal{S})^{1/2}\} \leq \frac{1}{2} c\nu p(\mathcal{S})$. Hence, if $C_7 = C_7(A, B)$ is chosen sufficiently large,

$$P\Big\{\tfrac{1}{2} c\nu\, p(\mathcal{S}) \leq M(\mathcal{S}) \leq 2c\nu\, p(\mathcal{S}) \text{ for all discs } \mathcal{S} \in \mathcal{D}_1(C_7)\Big\} \to 1\,. \tag{6.16}$$

Divide $\mathcal{R}$ into a lattice of squares of edge width $(\nu^{-1} \log \nu)^{(1+\epsilon)/(2-\alpha)}$, for some $\epsilon > 0$. Given $C_8 > 1$, at the centre of any square Q in the lattice place the centre of a disc $\mathcal{S}(Q)$ for which $p\{\mathcal{S}(Q)\} = C_7 C_8 \nu^{-1} \log \nu$. For each $C_8 > 1$ the following is true: for all sufficiently large ν, and all Q, each disc contained in $\mathcal{D}_1(C_7)$ and centred within Q is a subset of $\mathcal{S}(Q)$. For any given ν, let $\mathcal{D}_2$ denote the set of discs $\mathcal{S}(Q)$, with Q ranging over all squares in the lattice such that $\mathcal{S}(Q) \subseteq \mathcal{R}$. Using (6.13) and (6.15) we may prove that for any $C_8, C_9 > 0$, there exists $C_3 > 0$, chosen sufficiently large, such that if $m \geq C_3 \log \nu$,

$$\sup_{\mathcal{S} \in \mathcal{D}_2} P\Big\{M(\mathcal{S}) \geq m \text{ or } 2cC_2^{-1}\nu\, p(\mathcal{S}) \geq m\Big\} = O\big(\nu^{-C_9}\big)\,.$$

Noting that $\mathcal{D}_2$ contains at most $O(\nu^{2(1+\epsilon)/(2-\alpha)})$ elements, we see that if C_7 (and hence C_3) is sufficiently large,

$$P\Big\{M(\mathcal{S}) \geq m \text{ or } 2cC_2^{-1}\nu\, p(\mathcal{S}) \geq m \text{ for some } \mathcal{S} \in \mathcal{D}_2\Big\} \to 0\,.$$

Therefore, by definition of $\mathcal{D}_2$,

$$P\Big\{M(\mathcal{S}) \geq m \text{ or } 2cC_2^{-1}\nu\, p(\mathcal{S}) \geq m \\ \text{for some } \mathcal{S} \subseteq \mathcal{R}' \text{ with } \mathcal{S} \notin \mathcal{D}_1(C_7)\Big\} \to 0\,.$$

Combining this result and (6.16) we deduce that for $C_3 > 0$ sufficiently large,

$$P\Big\{\tfrac{1}{2}c\nu\, p(\mathcal{S}) \le M(\mathcal{S}) \le 2c\nu\, p(\mathcal{S}) \text{ for all discs } \mathcal{S} \subseteq \mathcal{R}' \text{ with either } M(\mathcal{S}) \ge m \text{ or } 2cC_2^{-1}\nu\, p(\mathcal{S}) \ge m\Big\} \to 1\,.$$

Hence, defining $q(\mathcal{S}) = c^{-1}\int_{\mathcal{S}} \|x-v\|^{-\alpha}\,dx$ and taking C_3 larger if necessary, we see that with $C_{10} = \frac{1}{2}cC_1$ and $C_{11} = 2cC_2$, where C_1, C_2 are as in (6.9), $C_{10}\,q(\mathcal{S}) \le \frac{1}{2}\int_{\mathcal{S}} \|x-v\|^{-\alpha}\{\lambda(x)\,\|x-v\|^{\alpha}\}\,dx \le \frac{1}{2}c\,p(\mathcal{S})$ for all $\mathcal{S} \subseteq \mathcal{R}'$. Similarly, $C_{11}\,q(\mathcal{S}) \ge 2c\,p(\mathcal{S})$. Therefore, we have

$$P\Big\{C_{10}\nu\, q(\mathcal{S}) \le M(\mathcal{S}) \le C_{11}\nu\, q(\mathcal{S}) \text{ for all discs } \mathcal{S} \subseteq \mathcal{R}' \text{ with either } M(\mathcal{S}) \ge m \text{ or } 2c\nu\, q(\mathcal{S}) \ge m\Big\} \to 1\,.$$

Let $\mathcal{S}_0$ denote the disc centred at v and of radius r_1, the latter defined by $\frac{1}{4}cC_1\nu\, q(\mathcal{S}_0) = m$. Then,

$$p(\mathcal{S}_0) \ge C_1\, q(\mathcal{S}_0) \ge C_1\big(\tfrac{1}{4}cC_1\nu\big)^{-1} C_3 \log\nu \ge \nu^{-1}\log\nu\,,$$

provided $C_3 \ge c/4$. In this case, (6.14) is satisfied by $\mathcal{S}_0$. Also, with B as in (6.15),

$$\tfrac{1}{4}c\nu\, p(\mathcal{S}_0)\big/\big[B\,\{\nu\, p(\mathcal{S}_0)\,\log\nu\}^{1/2}\big] \ge \tfrac{1}{4}cB^{-1}\big\{C_1\,(\tfrac{1}{4}cC_1)^{-1}C_3\big\}^{1/2} \ge 1$$

if $C_3 \ge 4B^2/c$. Choose C_3 so large that it satisfies both these conditions. Then, by (6.13) and (6.15),

$$\begin{aligned} P\{M(\mathcal{S}_0) \ge m\} &\ge P\Big\{M(\mathcal{S}_0) - N\,p(\mathcal{S}_0) \ge \tfrac{1}{4}cC_1\nu\, q(\mathcal{S}_0) - \tfrac{1}{2}c\nu\, p(\mathcal{S}_0)\Big\} + o(1) \\ &\ge P\Big\{M(\mathcal{S}_0) - N\,p(\mathcal{S}_0) \ge -\tfrac{1}{4}c\nu\, p(\mathcal{S}_0)\Big\} + o(1) \to 1\,. \end{aligned}$$

Let $\mathcal{D}_3$ be the class of all discs which are contained within $\mathcal{R}'$ and whose centres are distant at least $C_{12}\,r_1$ from v, where $C_{12} > 0$ is a constant. If C_{12} is chosen sufficiently large, depending on α, c, C_1, C_{11}, then any $\mathcal{S} \in \mathcal{D}_3$ which satisfies $C_{11}\nu\, q(\mathcal{S}) \ge m$, i.e. which satisfies $C_{11}\,q(\mathcal{S}) \ge \frac{1}{4}cC_1\,q(\mathcal{S}_0) = \frac{1}{2}C_{10}\,q(\mathcal{S}_0)$, has radius exceeding r_1, and so has larger area than $\mathcal{S}_0$. If $C_{10}\nu\, q(\mathcal{S}) \le M(\mathcal{S}) \le C_{11}\nu\, q(\mathcal{S})$ and $M(\mathcal{S}) \ge m$ then $C_{11}\nu\, q(\mathcal{S}) \ge m$.

Therefore, with probability tending to 1, (a) $M(\mathcal{S}_0) \ge m$, and (b) any disc $\mathcal{S}$ whose centre is distant at least $C_{12}\,r_1$ from v and which satisfies $M(\mathcal{S}) \ge m$ has larger area than $\mathcal{S}_0$. Hence, with probability tending to 1, the disc $\widehat{\mathcal{S}} \subseteq \mathcal{R}'$ with

smallest area subject to $M(\widehat{\mathcal{S}}) \geq m$, has its centre distant less than $C_{12}\, r_1$ from v. That is, $P(\|\hat{v} - v\| \leq C_{12}\, r_1) \to 1$. Since $r_1 = C_4\, C_{12}^{-1}\, (m/\nu)^{1/(2-\alpha)}$ where $C_4\, C_{12}^{-1} = (\frac{1}{2}\pi c C_1)^{1/(2-\alpha)}$, the theorem is proved.

Next we investigate the estimator $\hat{\alpha}$ defined at (6.4). Write $\widehat{\mathcal{S}}$ for the annulus $\mathcal{S}_2 \backslash \mathcal{S}_1$, centred at $\hat{v}$. Let its outer radius be u and its inner radius θu, where $0 < \theta < 1$. Replacing $\mathcal{R}$ by $\widehat{\mathcal{S}}$ and v by $\hat{v}$ in (6.3), and setting the right-hand side of (6.3) to 0, $\hat{\alpha}$ may be defined as the unique solution of the following equation in ξ:

$$\frac{\sum_i \left(\log \|X_i - \hat{v}\|\right) I(X_i \in \widehat{\mathcal{S}})}{\sum_i I(X_i \in \widehat{\mathcal{S}})} = \frac{\int_{\mathcal{S}} \left(\log \|x - \hat{v}\|\right) \|x - \hat{v}\|^{-\xi}\, dx}{\int_{\mathcal{S}} \|x - \hat{v}\|^{-\xi}\, dx}\,.$$

In Theorems 6.2 and 6.3 below we assume that the parameter θ, used to define $\hat{\alpha}$, $\tilde{\alpha}$ and $\check{\alpha}$, is fixed in the range $0 < \theta < 1$, although there are versions of both those results for θ decreasing to zero sufficiently slowly. In place of (6.9) we suppose that

there exist constants $0 < \beta < \alpha < 2$, and functions a, b, such that a has two bounded derivatives, $a(v) > 0$, b is continuous at v, and

$$\left|\lambda(x) - \left\{a(x)\, \|x - v\|^{-\alpha} + b(x)\, \|x - v\|^{-\beta}\right\}\right| = o\left(\|x - v\|^{-\beta}\right) \qquad (6.17)$$

as $x \to v$.

Theorem 6.2 *Assume condition (6.17), and that $u = u(\nu) \to 0$ such that $\Delta \equiv \|\hat{v} - v\|/u \to 0$ in probability, that $\nu^{1-\epsilon} u^{2-\alpha} \to \infty$ for some $\epsilon > 0$, and that $(\nu^{1-\epsilon} u^{2-\alpha})^{-1} \Delta = o_p(1)$. Then, there exist constants $c_1 \neq 0$ and $c_2 > 0$ such that*

$$\begin{aligned} \hat{\alpha} = \alpha &+ c_1\, b(v)\, |\log u|^{-1}\, u^{\alpha-\beta} + c_2\, (\log u)^{-2} \left(\nu u^{2-\alpha}\right)^{-1/2} Z \\ &+ O_p\left(|\log u|^{-1}\, \Delta\right) + o_p\left(|\log u|^{-1}\, u^{\alpha-\beta}\right) \end{aligned} \qquad (6.18)$$

as $\nu \to \infty$, where Z is asymptotically Normal $N(0,1)$.

Remark 6.2. *Rate of convergence of $\hat{\alpha}$ to α.* Neglecting for the time being the term in $|\log u|^{-1}\Delta$, the optimal rate of convergence of $\hat{\alpha}$ to α is obtained by equating the orders of the second and third terms on the right-hand side of (6.18). This suggests that ideally, the smoothing parameter u should be taken equal to a constant multiple of $u_1 \equiv \{\nu (\log \nu)^2\}^{-1/(\alpha-2\beta+2)}$, in which case we have by (6.18) that

$$\hat{\alpha} - \alpha = O_p\left\{(\log \nu)^{-(3\alpha-4\beta+2)/(\alpha-2\beta+2)}\, \nu^{-(\alpha-\beta)/(\alpha-2\beta+2)} + (\log \nu)^{-1}\, \Delta\right\}. \qquad (6.19)$$

We showed in Remark 6.1 that we may estimate v at rate $\delta \equiv (\nu^{-1}\log\nu)^{1/(2-\alpha)}$, under conditions weaker than those imposed in Theorem 6.2. Therefore, we can in principle construct $\hat{v}$ so that Δ is of order $u_1^{-1}\delta$, and hence so that

$$(\log\nu)^{-1}\Delta = O_p\left\{(\log\nu)^{\{(\alpha-1)(\alpha-2\beta)+2\}/(\alpha-2\beta+2)(2-\alpha)}\nu^{-2(\alpha-\beta)/\{(\alpha-2\beta+2)(2-\alpha)\}}\right\},$$

which is of smaller order than the first term within braces in (6.19). In this case, (6.19) reduces to

$$\hat{\alpha}-\alpha = O_p\left\{(\log\nu)^{-(3\alpha+2)/(\alpha-2\beta+2)}\nu^{-(\alpha-\beta)/(\alpha-2\beta+2)}\right\}. \tag{6.20}$$

Proof of Theorem 6.2. We assume for the sake of simplicity that a, b are constant functions. Without loss of generality, $v = 0$. Then, for all $x \in \widehat{\mathcal{S}}$, where $\widehat{\mathcal{S}}$ is the annulus with radii θu and u centred at $\hat{v}$,

$$\left|\|x\|\,\|x-\hat{v}\|^{-1}-1\right| \le \|\hat{v}\|\,\|x-\hat{v}\|^{-1} \le \|\hat{v}\|\,(\theta u)^{-1} = \theta^{-1}\Delta,$$

and so,

$$\begin{aligned}
&\sum_i (\log\|X_i-\hat{v}\|)\, I(X_i\in\widehat{\mathcal{S}}) \\
&\quad\le \sum_i \log\left\{(1+\Delta)^{-1}\|X_i\|\right\} I(X_i\in\widehat{\mathcal{S}}) \\
&\quad= \sum_i (\log\|X_i\|)\, I(X_i\in\widehat{\mathcal{S}}) + O_p\Big\{\Delta\sum_i I(X_i\in\widehat{\mathcal{S}})\Big\}.
\end{aligned}$$

Let $\mathcal{S}$ denote the annulus with radii $\theta u, u$ centred at $v = 0$. For $t = \theta u$ or u, let $\mathcal{T}(t) = \{x : (1-\Delta)t \le \|x\| \le (1+\Delta)t\}$. Since $x \in \widehat{\mathcal{S}}\cup\mathcal{S}$ implies $x \in \mathcal{T}(\theta u)\cup\mathcal{T}(u)$, and $\sum_i I\{X_i\in\mathcal{T}(u)\} = O_p(\nu u^{2-\alpha}\Delta)$, then

$$\begin{aligned}
&\Big|\sum_i (\log\|X_i\|)\, I(X_i\in\widehat{\mathcal{S}}) - \sum_i (\log\|X_i\|)\, I(X_i\in\mathcal{S})\Big| \\
&\le \sum_i \big|\log\|X_i\|\big|\, I(X_i\in\widehat{\mathcal{S}}\cup\mathcal{S}) \\
&\le \sum_i \big|\log\|X_i\|\big|\, I\{X_i\in\mathcal{T}(\theta u)\cup\mathcal{T}(u)\} \\
&= O_p\Big[|\log u_1|\sum_i I\{X_i\in\mathcal{T}(\theta u)\} + |\log u|\sum_i I\{X_i\in\mathcal{T}(u)\}\Big] \\
&= O_p\{|\log u|\,(\nu u^{2-\alpha}\Delta+\nu^{\epsilon})\}
\end{aligned}$$

for all $\epsilon > 0$. Similarly,

$$\begin{aligned} \Big| \sum_i I(X_i \in \widehat{\mathcal{S}}) - \sum_i I(X_i \in \mathcal{S}) \Big| &= O_p(\nu u^{2-\alpha} \Delta + \nu^\epsilon) , \\ \sum_i I(X_i \in \mathcal{S}) &= \{1 + o_p(1)\} \text{const.} \, \nu u^{2-\alpha} \end{aligned}$$

for all $\epsilon > 0$, where const. denotes a positive constant. Therefore, since $\nu^{1-\epsilon} u^{2-\alpha} \to \infty$ for some $\epsilon > 0$,

$$\begin{aligned} &\frac{\sum_i (\log \|X_i - \hat{v}\|) \, I(X_i \in \widehat{\mathcal{S}})}{\sum_i I(X_i \in \widehat{\mathcal{S}})} - \frac{\sum_i (\log \|X_i\|) \, I(X_i \in \mathcal{S})}{\sum_i I(X_i \in \mathcal{S})} \\ &= \frac{\sum_i (\log \|X_i\|) \, I(X_i \in \mathcal{S}) + O_p\big\{ |\log u| \, (\nu u^{2-\alpha} \Delta) + (|\log u| + \Delta) \, \nu^\epsilon \big\}}{\sum_i I(X_i \in \mathcal{S}) + O_p(\nu u^{2-\alpha} \Delta + \nu^\epsilon)} \\ &\qquad - \frac{\sum_i (\log \|X_i\|) \, I(X_i \in \mathcal{S})}{\sum_i I(X_i \in \mathcal{S})} \\ &= O_p(\Delta |\log u|) . \end{aligned} \tag{6.21}$$

An argument similar to that used to derive (6.32) (but simpler, since now — after conditioning — we have two sums of independent random variables, not a U-statistic) may be employed to prove that

$$\begin{aligned} \frac{\sum_i (\log \|X_i\|) \, I(X_i \in \mathcal{S})}{\sum_i I(X_i \in \mathcal{S})} - \frac{\int_{\mathcal{S}} (\log \|x\|) \, \lambda(x) \, dx}{\int_{\mathcal{S}} \lambda(x) \, dx} \\ = (\nu u^{2-\alpha})^{-1/2} \, c_3 \, Z_1 , \end{aligned} \tag{6.22}$$

where $c_3 > 0$ and Z_1 has an asymptotic Normal $N(0,1)$ distribution.

Let η_u denote any sequence of positive numbers converging to zero more rapidly than $|\log u|^{-1}$. By Taylor expansion, for $i = 0$ or 1,

$$\begin{aligned} \int_{\mathcal{S}} (\log \|x\|)^i \|x\|^{-\xi} \, dx &= \int_{\mathcal{S}} (\log \|x\|)^i \|x\|^{-\alpha} \, dx \\ &\quad + (\alpha - \xi) \int_{\mathcal{S}} (\log \|x\|)^{i+1} \|x\|^{-\alpha} \, dx \\ &\quad + O\big\{ (\alpha - \xi)^2 |\log u|^{i+2} \, u^{2-\alpha} \big\} , \end{aligned} \tag{6.23}$$

uniformly in $\xi \in (\alpha - \eta_u, \alpha + \eta_u)$, as $u \to 0$; and

$$\int_{\mathcal{S}} (\log \|x\|)^i \, \lambda(x) \, dx = a \int_{\mathcal{S}} (\log \|x\|)^i \|x\|^{-\alpha} \, dx$$

$$+b\int_{\mathcal{S}}(\log\|x\|)^i\|x\|^{-\beta}\,dx$$
$$+o\big(|\log u|^i\,u^{1-\beta}\big)\,, \tag{6.24}$$

$$\int_{\mathcal{S}}(\log\|x\|)^i\|x\|^{-\alpha}\,dx \;\sim\; c_4(\log u)^i\,u^{2-\alpha}\,, \tag{6.25}$$

where $c_4 = c_4(i) > 0$. (In the case where a, b are functions rather than constants, and when deriving (6.25) and replacing a, b on the right-hand side by $a(v)$, $b(v)$ respectively, we require up to two derivatives of $a(x)$ and continuity of $b(x)$. In the event that $\beta \le \alpha - 1$ there might appear to be an additional contribution to bias, arising from the first derivative of $a(x)$, but in fact it vanishes.) From the last two results we deduce that

$$\frac{\int_{\mathcal{S}}(\log\|x\|)\,\lambda(x)\,dx}{\int_{\mathcal{S}}\lambda(x)\,dx} = \frac{\int_{\mathcal{S}}(\log\|x\|)\,\|x\|^{-\alpha}\,dx}{\int_{\mathcal{S}}\|x\|^{-\alpha}dx}$$
$$+(b/a)\,c_5(\log u)\,u^{\alpha-\beta} + o\big(|\log u|u^{\alpha-\beta}\big)\,, \tag{6.26}$$

where $c_5 > 0$.

Combining (6.21)–(6.23), (6.25) and (6.26) we obtain,

$$\frac{\sum_i(\log\|X_i-\hat v\|)\,I(X_i\in\widehat{\mathcal{S}})}{\sum_i I(X_i\in\widehat{\mathcal{S}})} - \frac{\int_{\mathcal{S}}(\log\|x\|)\,\|x\|^{-\xi}\,dx}{\int_{\mathcal{S}}\|x\|^{-\xi}\,dx}$$
$$= (\nu u^{2-\alpha})^{-1/2}c_3Z_1 + (b/a)\,c_5(\log u)\,u^{\alpha-\beta} - (\alpha-\xi)\,c_6\,(\log u)^2$$
$$+o_p\Big\{(\nu u^{2-\alpha})^{-1/2} + |\log u|\,u^{\alpha-\beta} + |\alpha-\xi|\,(\log u)^2\Big\} + O_p(\Delta|\log u|)$$

uniformly in $\xi \in (\alpha-\eta_u, \alpha+\eta_u)$. Since the left-hand side of this formula is monotone in ξ, it can have at most one zero. Therefore, the formula proves that a zero $\xi = \hat\alpha$ exists with probability tending to 1, and satisfies $P\{\hat\alpha \in (\alpha-\eta_u, \alpha+\eta_u)\} \to 1$ if $\eta_u \to 0$ sufficiently slowly (but faster than $1/(|\log u|)$; and

$$(\alpha-\hat\alpha)c_6(\log u)^2 = (\nu u^{2-\alpha})^{-1/2}\,c_3\,Z_1 + (b/a)\,c_5\,(\log u)u^{\alpha-\beta}$$
$$+o_p\Big\{\big(\nu u^{2-\alpha}\big)^{-1/2} + |\log u|\,u^{\alpha-\beta}\Big\} + O_p(\Delta|\log u|)\,.$$

The theorem follows from this result.

Finally we describe properties of the estimators $\tilde\alpha$ and $\check\alpha$, introduced in Section 6.4. Let $\bar\alpha$ denote either estimator.

Theorem 6.3 *Assume the conditions of Theorem 6.2, except that in (6.17) the assumption* $0 < \beta < \alpha < 2$ *should be strengthened to* $1 < \beta < \alpha < 2$, *while the function* $a(\cdot)$ *need have only one bounded derivative. Then, there exist constants* $c_1 \neq 0$ *and* $c_2 > 0$ *such that,*

$$\bar{\alpha} = \alpha + c_1 b(v)\, u^{\alpha-\beta} + c_2 \left(\nu u^{2-\alpha}\right)^{-1/2} Z + o_p\left(u^{\alpha-\beta}\right) \tag{6.27}$$

as $\nu \to \infty$, *where* Z *is asymptotically Normal* $N(0,1)$. *(The constants* c_1 *and* c_2 *assume different values in the cases* $\bar{\alpha} = \tilde{\alpha}$, $\bar{\alpha} = \check{\alpha}$.*)*

Remark 6.3. *Rate of convergence of* $\bar{\alpha}$ *to* α. Arguing as in Remark 6.2, we may show that by equating the orders of the second and third terms on the right-hand side of (6.27), the optimal rate of convergence of $\bar{\alpha}$ to α is

$$\bar{\alpha} - \alpha = O_p\left(\nu^{-(\alpha-\beta)/(\alpha-2\beta+2)}\right), \tag{6.28}$$

and is achieved with u equal to a constant multiple of $\nu^{-1/(\alpha-2\beta+2)}$. This rate is a little slower than that of $\hat{\alpha}$ to α, given in (6.20).

Remark 6.4. *Inferiority of the estimators* $\tilde{\alpha}$ *and* $\check{\alpha}$. When $1 < \alpha < 2$, each of the estimators $\hat{\alpha}, \tilde{\alpha}, \check{\alpha}$ is consistent. We claim, however, that for such α's, and when the simplistic model $\lambda(x) = c\,\|x - v\|^{-\alpha}$ is exactly correct (or valid to a high degree of accuracy), $\hat{\alpha}$ can be constructed so that it enjoys substantially greater accuracy than any construction of either $\tilde{\alpha}$ or $\check{\alpha}$. This good performance of $\hat{\alpha}$ goes well beyond the advantages conferred by the logarithmic factor in (6.20), compared to (6.28). To appreciate why, note that the restriction $1 < \beta < \alpha < 2$ in Theorem 6.3 prevents us from taking β arbitrarily close to 0, which we would do in the case of the simplistic model. On the other hand, arbitrarily small β's are possible for Theorem 6.2; this allows us to achieve much smaller biases for the estimator $\hat{\alpha}$ there. The restriction $1 < \beta < \alpha$ is unfortunately essential to Theorem 6.3. Indeed, if $\beta < 1$ then the term $c_1 b(v)\, u^{\alpha-\beta}$ on the right-hand side of (6.27) should be replaced by $c_3\, u^{\alpha-1}$, where c_3 is a nonzero constant depending globally on λ, not just on its behaviour in the neighbourhood of v. This prevents the estimators $\tilde{\alpha}$ and $\check{\alpha}$ from enjoying the low bias, and consequently fast rate of convergence, of $\hat{\alpha}$ under the simplistic model.

Proof of Theorem 6.3. We consider only the case $\bar{\alpha} = \tilde{\alpha}$, since that of $\check{\alpha}$ is similar. Under the conditions of the theorem, $\alpha - \beta < 1$ and the function a has

a bounded derivative. Hence, there is no loss of generality in assuming that it is constant. Furthermore, it suffices to establish the theorem for $\tilde{\gamma}$:

$$\tilde{\gamma} = \gamma + c_1 b(v) u^{(\alpha-\beta)} + c_2 \big(\nu u^{2-\alpha}\big)^{-1/2} Z + o_p\big(u^{(\alpha-\beta)}\big)\,, \tag{6.29}$$

where $\gamma = \{2(2-\alpha)\}^{-1}$ and c_1, c_2 and Z have the stated properties of, while being different from, those quantities in the theorem. Let X, Y_1, Y_2 denote independent random variables with density $f = \lambda / \int_{\mathcal{R}} \lambda$ on $\mathcal{R}$, and put $n = \nu \int_{\mathcal{R}} \lambda$ (not necessarily an integer),

$$\begin{aligned}
p &= P(\|Y_1 - Y_2\| \le u) = E\{I(\|Y_1 - Y_2\| \le u)\}\,,\\
q_1 &= E(\log \|Y_1 - Y_2\| \,|\, \|Y_1 - Y_2\| \le u)\,,\\
q &= (\log u - q_1)\, p = \int_0^u t^{-1} P(\|Y_1 - Y_2\| \le t)\, dt\,,\\
\Delta_1 &= \Big\{ \sum\sum_{i<j} I(\|X_i - X_j\| \le u) - \tfrac{1}{2} n^2 p \Big\} \Big/ \big(\tfrac{1}{2} n^2 p\big)\,,\\
\Delta_2 &= \Big\{ \sum\sum_{i<j} (\log \|X_i - X_j\| - q_1)\, I(\|X_i - X_j\| \le u) \Big\} \Big/ \big(\tfrac{1}{2} n^2 p\big)\,.
\end{aligned}$$

Since $\operatorname{var}\{\sum\sum_{i<j} I(\|X_i - X_j\| \le u)\} = O(n^3)$ then $\Delta_1 = o_p(1)$. Moreover, using the above notation, we have

$$\begin{aligned}
\tilde{\gamma} &= \frac{\sum\sum_{i<j} (\log u - \log \|X_i - X_j\|)\, I(\|X_i - X_j\| \le u)}{\sum\sum_{i<j} I(\|X_i - X_j\| \le u)}\\
&= \frac{\sum\sum_{i<j} (\log u - q_1 + q_1 - \log \|X_i - X_j\|)\, I(\|X_i - X_j\| \le u)}{\sum\sum_{i<j} I(\|X_i - X_j\| \le u)}\\
&= (\log u - q_1) - \frac{\sum\sum_{i<j} (\log \|X_i - X_j\| - q_1)\, I(\|X_i - X_j\| \le u)}{\frac{1}{2} n^2 p}\\
&\qquad \times \frac{\frac{1}{2} n^2 p}{\frac{1}{2} n^2 p + \sum\sum_{i<j} I(\|X_i - X_j\| \le u) - \frac{1}{2} n^2 p}\\
&= p^{-1} q - \Delta_2 (1 + \Delta_1)^{-1}\\
&= p^{-1} q - \{1 + o_p(1)\}\, \Delta_2\,.
\end{aligned}$$

Therefore, (6.29) will follow if we prove that

$$p \sim \text{const.}\, u^{2(2-\alpha)}\,, \tag{6.30}$$

$$p^{-1} q = \gamma + c_1 b(v) u^{\alpha-\beta} + o\big(u^{\alpha-\beta}\big)\,, \tag{6.31}$$

$$\tfrac{1}{2} n^2 p \Delta_2 = \text{const.}\, \big(\nu u^{2-\alpha}\big)^{3/2} Z_1\,, \tag{6.32}$$

where the quantities denoted by "const." are strictly positive constants, and Z_1 is asymptotically Normal $N(0,1)$.

Put

$$Q(u,x) = \int_{0<t<u} t^{-1} P(\|x - X\| \le t)\, dt - \gamma P(\|x - X\| \le u)\,.$$

Lemma 6.3. *Assume condition (6.17), except that $0 < \beta < \alpha < 2$ should be strengthened to $1 < \beta < \alpha < 2$. Then, for all sufficiently small $\delta > 0$, and as $u \to 0$,*

$$E\{|Q(u,X)|^{2+\delta}\} = O(u^{(2-\alpha)(3+\delta)})\,, \tag{6.33}$$

$$E\{Q(u,X)^2\} \sim c_3 u^{3(2-\alpha)}\,, \tag{6.34}$$

and (6.30) and (6.31) hold, where $c_3 > 0$.

To derive (6.32) from the lemma, let $K = K(\mathcal{R})$ be the number of Poisson points X_i that lie in $\mathcal{R}$; denote these by $X_1, \ldots, X_K$. Conditional on K they are independent random variables distributed as X (see Section 6.2). Put $h(x,y) = (\log \|x - y\| - q_1)\, I(\|x - y\| \le u)$ and $g(x) = E\{h(x,X)\}$. We treat Δ_2 as a second-order U-statistic, writing

$$\tfrac{1}{2} n^2 p \Delta_2 = \sum\sum_{1 \le i < j \le K} h(X_i, X_j)\,.$$

Putting $T_1 = \sum_{i=1}^K g(X_i)$ and $T_2 = \sum\sum_{1 \le i < j \le K} \{h(X_i,X_j) - g(X_i) - g(X_j)\}$, we see that

$$\begin{aligned}
\tfrac{1}{2} n^2 p \Delta_2 &= \sum\sum_{1 \le i < j \le K} \{h(X_i,X_j) - g(X_i) - g(X_j)\} + \sum\sum_{1 \le i < j \le K} \{g(X_i) + g(X_j)\} \\
&= T_2 + \sum_{i=2}^K (K - i) g(X_i) + \sum_{j=2}^K (j-1) g(X_j) \\
&= T_2 + (K-1) \sum_{i=1}^K g(X_i) = T_2 + (K-1) T_1\,.
\end{aligned} \tag{6.35}$$

Note that $K = \{1 + o_p(1)\} n = \{1 + o_p(1)\} \nu \int_{\mathcal{R}} \lambda$. Therefore, a central limit theorem for T_1, of the form $\{T_1 - KEg(X)\}/\{K \operatorname{var} g(X)\}^{1/2} \to N(0,1)$ in distribution, will follow from Lyapounov's theorem (see for example, Chung, 1974, Chapter 7) if we prove that for some positive δ,

$$\nu E\{g(X)^2\} \to \infty\,, \quad \frac{E\{|g(X)|^{2+\delta}\}}{\nu^{\delta/2} \{\operatorname{var} g(X)\}^{1+(\delta/2)}} \to 0\,. \tag{6.36}$$

Observe too that $E\{g(X)\} = 0$. Since

$$\begin{aligned} E(T_2^2|K) &= \tfrac{1}{2}K(K-1)\,E\{h(Y_1,Y_2) - g(Y_1) - g(Y_2)\}^2 \\ &= O_p\big[\nu^2 E\big\{h(Y_1,Y_2)^2 + g(Y_1)^2\big\}\big]\,, \end{aligned}$$

then, assuming we have proved the central limit theorem for T_1, the contribution of T_2 to (6.35) will be asymptotically negligible if we show that

$$\text{var}\,(T_2|K) \,/\, \text{var}\,\{(K-1)T_1|K\} \to 0\,,$$

which is equivalent to proving that

$$E\big\{h(Y_1,Y_2)^2\big\}/\big\{\nu E g(X)^2\big\} \to 0\,. \tag{6.37}$$

Now,

$$\begin{aligned} &E\{(\log\|Y_1 - Y_2\| - \log u)^2\, I(\|Y_1 - Y_2\| \le u)\} \\ &\quad = \int_0^u (\log t - \log u)^2\, dP(\|Y_1 - Y_2\| \le t) \\ &\quad = O\bigg\{\int_0^u (\log u - \log t)\, t^{2(2-\alpha)-1}\, dt\bigg\} \\ &\quad = O\big(u^{2(2-\alpha)}\big)\,, \end{aligned}$$

using (6.30) and integration by parts. Note too that $(q_1 - \log u)^2 p(u)$ has the same order, since

$$\begin{aligned} &\{E(\log\|Y_1 - Y_2\| \,|\, \|Y_1 - Y_2\| \le u) - \log u\}^2\, p(u) \\ &\quad = p(u)^{-1}\bigg\{\int_0^u (\log t - \log u)\, dP(\|Y_1 - Y_2\| \le t)\bigg\}^2 \\ &\quad = O(u^{2(2-\alpha)})\,. \end{aligned}$$

Therefore, recalling the definition of $h(Y_1, Y_2)$, and noting that

$$\begin{aligned} E\big\{h(Y_1,Y_2)^2\big\} &= E\big\{(\log\|Y_1 - Y_2\| - \log u)^2\, I(\|Y_1 - Y_2\| \le u) \\ &\qquad + (\log\|Y_1 - Y_2\| - \log u)(\log u - q_1)\, I(\|Y_1 - Y_2\| \le u) \\ &\qquad + (\log u - q_1)^2\, I(\|Y_1 - Y_2\| \le u)\big\}\,, \end{aligned}$$

it may be shown that $E\{h(Y_1,Y_2)^2\} = O(u^{2(2-\alpha)})$. (The middle term on the right-hand side may similarly be shown to be of order $u^{2(2-\alpha)}$.) Hence, (6.37) will follow if we prove that

$$p(u)/\big\{\nu E g(X)^2\big\} \to 0\,. \tag{6.38}$$

Define $g_1(x) = (\log u - q_1 - \gamma)\, P(\|x - X\| \le u)$ and

$$g_2(x) = E\{(\log u - \log \|x - X\|)\, I(\|x - X\| \le u)\} - \gamma P(\|x - X\| \le u) = Q(u, x)\,.$$

Then $g = g_1 - g_2$. In view of (6.31),

$$\begin{aligned} |\, g_1(x)\,| &= |\, (p^{-1}q - \gamma)\, P(\|x - X\| \le u)\,| \\ &\le C u^{\alpha-\beta}\, P(\|x - X\| \le u)\,, \end{aligned}$$

where C depends on neither u nor x. From this it may be proved that g_1 makes a negligible contribution to g in deriving (6.36) and (6.38). Therefore, it suffices to establish those results in the case where $g(x)$ is replaced by $Q(u, x)$; but there, they follow directly from (6.33) and (6.34), on recalling that $\nu u^{2-\alpha} \to \infty$:

$$\begin{aligned} \frac{E\{|Q(u,X)|^{2+\delta}\}}{\nu^{\delta/2}\{\operatorname{var} Q(u,X)\}^{(1+\delta/2)}} &= O\{(\nu u^{(2-\alpha)})^{-\delta/2}\}\,, \\ p(u)/\{\nu\, EQ(u,X)^2\} &= O\{(\nu u^{(2-\alpha)})^{-1}\}\,. \end{aligned}$$

This completes the proof of (6.29), and hence of the theorem.

Proof of Lemma 6.3. To simplify our argument we shall assume that the function b is Hölder continuous with exponent $\epsilon \in (0, \beta - 1)$. In this case, (6.17) may be written as

$$\left|\lambda(x) - \left\{a(v)\|x - v\|^{-\alpha} + b(v)\|x - v\|^{-\beta}\right\}\right| \le C_1 \|x - v\|^{\epsilon-\beta}\,, \tag{6.39}$$

where $C_1 > 0$. An additional argument, with more complex notation, is needed to treat the case where b is only continuous.

Write $x = (r\cos\theta, r\sin\theta)^T$ and $X = (R\cos\Theta, R\sin\Theta)^T$, where $0 \le r,\ R < \infty$ and $0 \le \theta,\ \Theta < 2\pi$; and put $V = \cos(\theta - \Theta)$. In this notation, $\|x - X\|^2 = r^2 + R^2 - 2rRV$, and so

$$\begin{aligned} P(\|x - X\| \le u) &= P\{(R - rV)^2 \le u^2 - r^2(1 - V^2)\} \\ &= P\big[rV - \{u^2 - r^2(1 - V^2)\}^{1/2} \le R \le rV \\ &\qquad + \{u^2 - r^2(1 - V^2)\}^{1/2}, u^2 - r^2(1 - V^2) \ge 0\big]\,. \end{aligned} \tag{6.40}$$

Define $f = \lambda/(\int_{\mathcal{R}} \lambda)$ and assume without loss of generality $v = 0$. Under the same conditions as in Lemma 6.3, there exist constants $a_1, C_2, s_0 > 0$ and b_1 such that

$$\left|f(x) - (a_1\|x\|^{-\alpha} + b_1\|x\|^{-\beta})\right| \le C_2\|x\|^{\epsilon-\beta} \tag{6.41}$$

for $\|x\| \le s_0$, and $f(x) = 0$ for $\|x\| > s_0$. (Indeed, $a_1 = a(v)/\int_{\mathcal{R}} \lambda$, $b_1 = b(v)/\int_{\mathcal{R}} \lambda$.) Put

$$c(\xi) = 1 \Big/ \int_0^{s_0} s^{1-\xi} ds\,, \qquad 0 < \xi < 2\,,$$

and let W, S_1, S_2, S_3 be independent random variables, W having the distribution of $\cos \Psi$ where Ψ is Uniformly distributed on $(0, 2\pi)$, and S_1, S_2, S_3 having respective densities $c(\alpha)\, s^{1-\alpha}$, $c(\beta)\, s^{1-\beta}$ and $c(\beta-\epsilon)\, s^{1+\epsilon-\beta}$ for $0 < s \le s_0$, where it is assumed that ϵ is so small that $\beta - \epsilon > 1$. Define $D = \{u^2 - r^2(1-W^2)\}^{1/2}$ and

$$\pi_i(u,r) = P\{rW - D \le S_i \le rW + D, u^2 - r^2(1-W^2) > 0\}\,.$$

In view of (6.41) we may choose a function ζ, with the property $|\zeta(x)| < 1$ for $x \in \mathbb{R}^2$, such that

$$f(x) = a_1 \|x\|^{-\alpha} + b_1 \|x\|^{-\beta} + C_2\, \zeta(x) \|x\|^{\epsilon-\beta}\,.$$

Put $d_1 = 2\pi a_1/c(\alpha)$, $d_2 = 2\pi b_1/c(\beta)$ and $d_3(x) = 2\pi C_2 \zeta(x)/c(\beta-\epsilon)$, where C_2 is as in (6.41). If we write $r = \|x\|$ and $R = \|X\|$, then it follows from (6.40) that for some suitable $\tilde{s} \in (0, s_0]$ with $d_3 = d_3(\tilde{s})$,

$$\begin{aligned} P(\|x - X\| \le u) &= P\{rW - D \le R \le rW + D, u^2 - r^2(1-W^2) \ge 0\} \\ &= \int_{0<y\le s_0} P\{|y - rW| \le D, u^2 - r^2(1-W^2) \ge 0\} \\ &\qquad \times 2\pi \{a_1 y^{1-\alpha} + b_1 y^{1-\beta} + C_2\, \zeta(y)\, y^{1+\epsilon-\beta}\}\, dy \\ &= \sum_{i=1}^{3} d_i \pi_i(u, \|x\|)\,. \end{aligned} \tag{6.42}$$

Let $\xi_1 = \alpha$, $\xi_2 = \beta$ and $\xi_3 = \beta - \epsilon$, and put $e_i = c(\xi_i)\,(2-\xi_i)^{-1} = s_0^{-(2-\xi_i)}$. Observe that $r < u(1-W^2)^{1/2}$ and that for $r \le u$, $rW - D < 0$. Then, with $W_0 = |W|$ and $U = u/(1-W_0^2)^{1/2}$, we have

$$\begin{aligned} \pi_i(u,r) &= P\{rW - D \le S_i \le rW + D,\ u^2 - r^2(1-W^2) > 0,\ r \le u\} \\ &\qquad + P\{rW - D \le S_i \le rW + D,\ u^2 - r^2(1-W^2) > 0,\ r > u\} \\ &= I(r \le u)\, P(S_i \le rW + D) \\ &\qquad + P(u < r \le U,\ W > 0,\ rW - D \le S_i \le rW + D) \\ &= \tfrac{1}{2}\big[\, I(r \le u)\{P(S_i \le rW_0 + D) + P(S_i \le -rW_0 + D)\} \\ &\qquad + P(u < r \le U,\ rW_0 - D \le S_i \le rW_0 + D)\,\big] \end{aligned}$$

$$= \tfrac{1}{2}\, e_i E\Big(I(r \le u) \sum_{j=1}^{2} \big[\min\{(-1)^j r W_0 + D, s_0\}\big]^{2-\xi_i}$$
$$+ I(u < r \le U)\big[\{\min(rW_0 + D, s_0)\}^{2-\xi_i} - \{\min(rW_0 - D, s_0)\}^{2-\xi_i}\big]\Big)$$
$$\le C_3\big[I(r \le u)\, u^{2-\xi_i} + E\{I(u < r \le U)(rW_0)^{1-\xi_i}\}u\big]\,, \tag{6.43}$$

since $D \le u$ for $0 \le r \le U$ and the leading terms in the expansion of $(rW_0 + D)^{2-\xi_i}$ and $(rW_0 - D)^{2-\xi_i}$ cancel each other. For $1 < \xi < 2$,

$$E\big[I\{u < r \le (1 - W_0^2)^{-1/2}u\}W_0^{1-\xi}\big]$$
$$\le C_4 \int_{\{1-(u/r)^2\}^{1/2}}^{1} w^{1-\xi}(1 - w^2)^{-1/2}\, dw$$
$$\le C_4\{1 - (u/r)^2\}^{(1-\xi)/2} \int_{\{1-(u/r)^2\}}^{1} (1 - w)^{-1/2}\, dw$$
$$\le C_5 \min(1, u/r)\,,$$

and so

$$\pi_i(u, r) \le C_6\{u^{2-\xi_i}\, I(r \le u) + u^2 r^{-\xi_i}\, I(r > u)\}\,. \tag{6.44}$$

By (6.42) and (6.44),

$$|Q(u,x)| = \Big| \int_{0<t<u} \sum_{i=1}^{3} (t^{-1} - \gamma u^{-1})\, d_i \pi_i(u, \|x\|)\, dt \Big|$$
$$\le C_7 \sum_{i=1}^{3} \{u^{2-\xi_i}\, I(\|x\| \le u) + u^2 \|x\|^{-\xi_i}\, I(\|x\| > u)\}\,,$$

and hence that for all sufficiently small $\delta \ge 0$,

$$E\{|Q(u,X)|^{2+\delta}\} \le C_8 \sum_{i=1}^{3} E\{|u^{2-\xi_i} I(\|X\| \le u)|^{2+\delta} + |u^2 \|X\|^{-\xi_i} I(\|X\| > u)|^{2+\delta}\}$$
$$= O(u^{(2-\alpha)(3+\delta)})\,,$$

whence follows (6.33).

Define $Q_1(u,x) = \int_{0<t<u} t^{-1} P(\|x - X\| \le t)\, dt$, $Q_2(u,x) = \gamma P(\|x - X\| \le u)$. In this notation, $Q = Q_1 - Q_2$, and so

$$E\{Q(u,X)^2\} = E\{Q_1(u,X)^2\} + E\{Q_2(u,X)^2\} - 2E\{Q_1(u,X)\, Q_2(u,X)\}\,.$$

We establish (6.34) by (a) proving asymptotic formulae for the three terms on the right-hand side, each of the form $\sim$ const. $u^{3(2-\alpha)}$; and (b) putting the formulae together and checking that the constants do not cancel. For brevity we shall outline only the most complex part of step (a), proving that as $u \to 0$,

$$E\{Q_1(u,X)^2\} \sim \text{const.}\, u^{3(2-\alpha)}\,. \tag{6.45}$$

Results (6.42) and (6.44) imply that for some $\delta > 0$,

$$\begin{aligned} E\{Q_1(u,X)^2\} &= E\Big\{\sum_{i,j=1}^{3} d_i d_j \pi_i(u,\|X\|)\pi_j(u,\|X\|)\Big\} \\ &= d_1^2\tau + O(u^{3(2-\alpha)+\delta})\,, \end{aligned} \tag{6.46}$$

where

$$\tau = \int_0^u \int_0^u (t_1 t_2)^{-1}\, E\{\pi_1(t_1,\|X\|)\,\pi_1(t_2,\|X\|)\}\, dt_1\, dt_2\,.$$

We may prove from (6.43) that

$$\begin{aligned} \tau \sim\ & \frac{1}{4}e_1^2 \int_0^u \int_0^u (t_1 t_2)^{-1}\, E\Big\{\prod_{i=1}^{2}\Big(I(S_1 \le t_i)\sum_{j=1}^{2}\big\{(-1)^j S_1 W_0 \\ & +(U_i^2 - S_1^2)^{1/2}(1-W_0^2)^{1/2}\big\}^{2-\alpha} \\ & +I(t_i < S_1 \le U_i)\big[\big\{S_1 W_0 + (U_i^2 - S_1^2)^{1/2}(1-W_0^2)^{1/2}\big\}^{2-\alpha} \\ & -\big\{S_1 W_0 - (U_i^2 - S_1^2)^{1/2}(1-W_0^2)^{1/2}\big\}^{2-\alpha}\big]\Big)\Big\}\, dt_1\, dt_2\,. \end{aligned}$$

On the right-hand side, replace S_1 by s and replace the expectation over W_0 and S by an expectation over W_0 and an integral in s over $0 < s < s_0$, against the element $c(\alpha)s^{1-\alpha}ds$. Now make the changes of variable $s = uz$ and $t_i = uy_i$, for $i = 1, 2$, where $0 < z < s_0/u$ and $0 < y_i < 1$. The indicator functions $I(S_1 \le t_i)$ and $I(t_i < S_1 \le U_i)$ change to $I(z \le y_i)$ and $I\{y_i < z \le y_i(1-W_0^2)^{-1/2}\}$, respectively; $S_1 W_0 \pm (U_i^2 - S_1^2)^{1/2}(1-W_0^2)^{1/2}$ changes to $u[zW_0 \pm \{y_i^2 - (1-W_0^2)z^2\}^{1/2}]$; and $s^{1-\alpha}ds$ changes to $u^{2-\alpha}z^{1-\alpha}dz$. Take $u^{3(2-\alpha)}$ outside the triple integral as a constant factor. Then, the integrand is nonnegative, the integral depends on u only through the upper bound s_0/u to z, and the infinite integral over z converges. Therefore, $\tau \sim$ const. $u^{3(2-\alpha)}$ as $u \to 0$, where the constant is positive. Result (6.34) follows from this formula and (6.46).

The argument leading to (6.42) also gives

$$p = p(u) = P(\|Y_1 - Y_2\| \le u) = \sum_{i,j=1}^{2} d_i d_j E\{\pi_i(u, S_j)\} + O\Big[\sum\sum_{ij}{}' E\{\pi_i(u, S_j)\}\Big]$$

as $u \to 0$, where $\sum\sum_{ij}{}'$ denotes summation over $1 \le i, j \le 3$ such that at least one of i, j equals 3. Using (6.43), and making changes of variable similar to those discussed in the previous paragraph, we may prove that if ϵ (in the definition $\xi_3 = \beta - \epsilon$) is sufficiently small,

$$E\{\pi_i(u, S_j)\} = C(i,j)\, u^{4-\xi_i-\xi_j} + O(u^{2(2-\xi_3)})$$

for $1 \le i, j \le 3$, where $C(i,j) = C(j,i)$ is a constant, $C(1,1) > 0$, and $C(1,2) \ne 0$ equals a nonzero constant multiplied by b_1 (see (6.41) for b_1). Hence,

$$p(u) = d_1^2\, C(1,1)\, u^{2(2-\alpha)} + 2d_1 d_2\, C(1,2)\, u^{4-\alpha-\beta} + O(u^{4-\alpha-\beta+\delta})$$

for some $\delta > 0$. Therefore, with $\gamma = \{2(2-\alpha)\}^{-1}$,

$$\begin{aligned} q = q(u) &= \int_0^u t^{-1} p(t)\, dt \\ &= d_1^2\, C(1,1)\gamma u^{2(2-\alpha)} + 2d_1 d_2\, C(1,2)\,(4-\alpha-\beta)^{-1}\, u^{4-\alpha-\beta} + O(u^{4-\alpha-\beta+\delta})\,. \end{aligned}$$

In consequence,

$$p^{-1}q = \gamma + 2d_1^{-1} d_2\, C(1,1)^{-1}\, C(1,2)\, \frac{(\beta-\alpha)\gamma}{4-\alpha-\beta}\, u^{\alpha-\beta} + O(u^{\alpha-\beta+\delta})\,.$$

This establishes (6.31).

6.8 Numerical Study

We first analysed simulated data, to assess the performance in finite samples of the methods proposed in Sections 6.3 and 6.4. The results are reported in Sections 6.8.1 and 6.8.2, and provide insight into methods appropriate for real data. With the benefit of this experience we applied our methods to data on seismic events in the vicinity of Tokyo. The results obtained are summarised in Section 6.8.3.

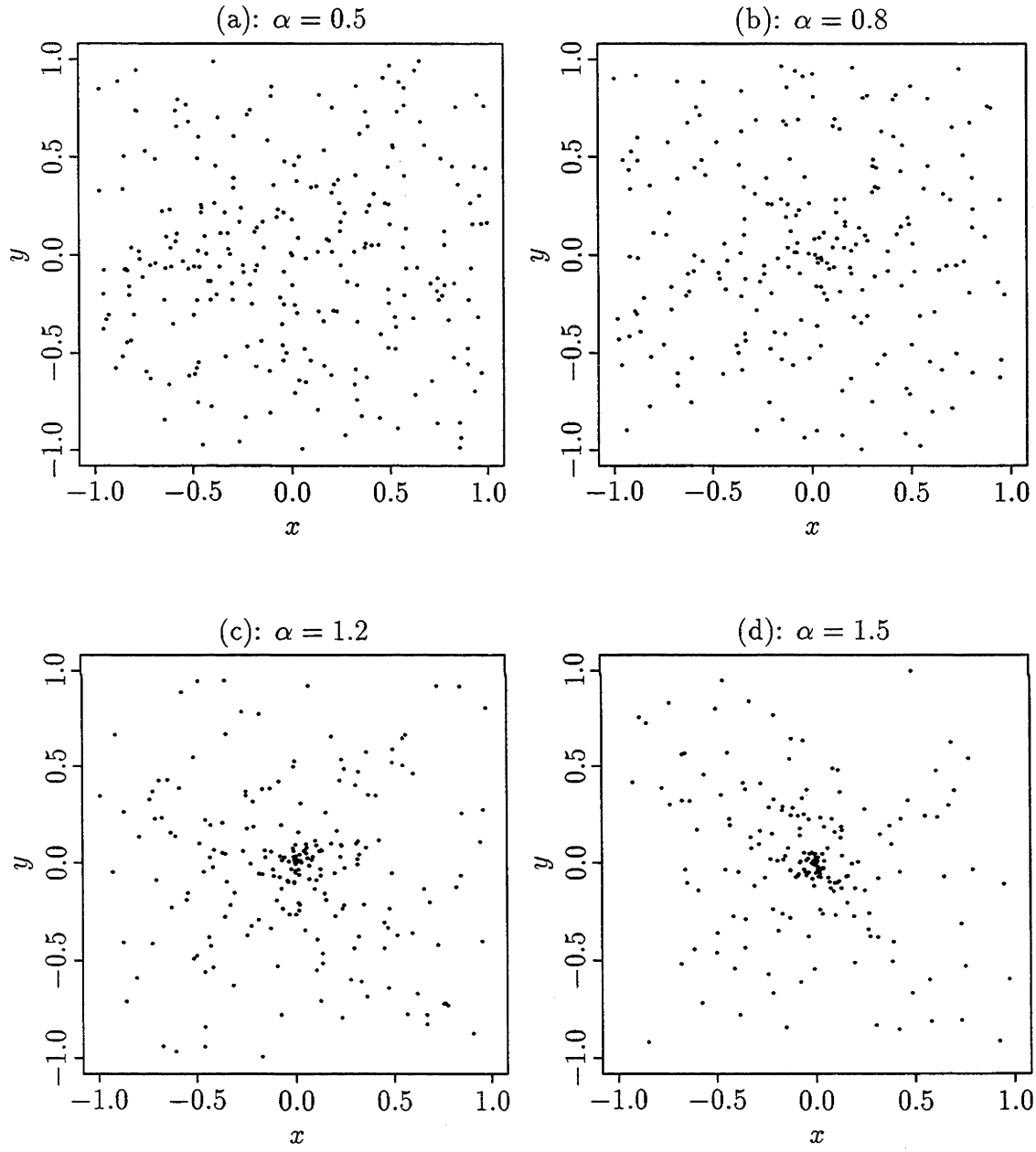


Figure 6.2: Simulated Poisson point process data on $[-1, 1]^2$. Plots in panels (a)–(d) correspond to pole strength $\alpha = 0.5, 0.8, 1.2$ and 1.5, respectively.

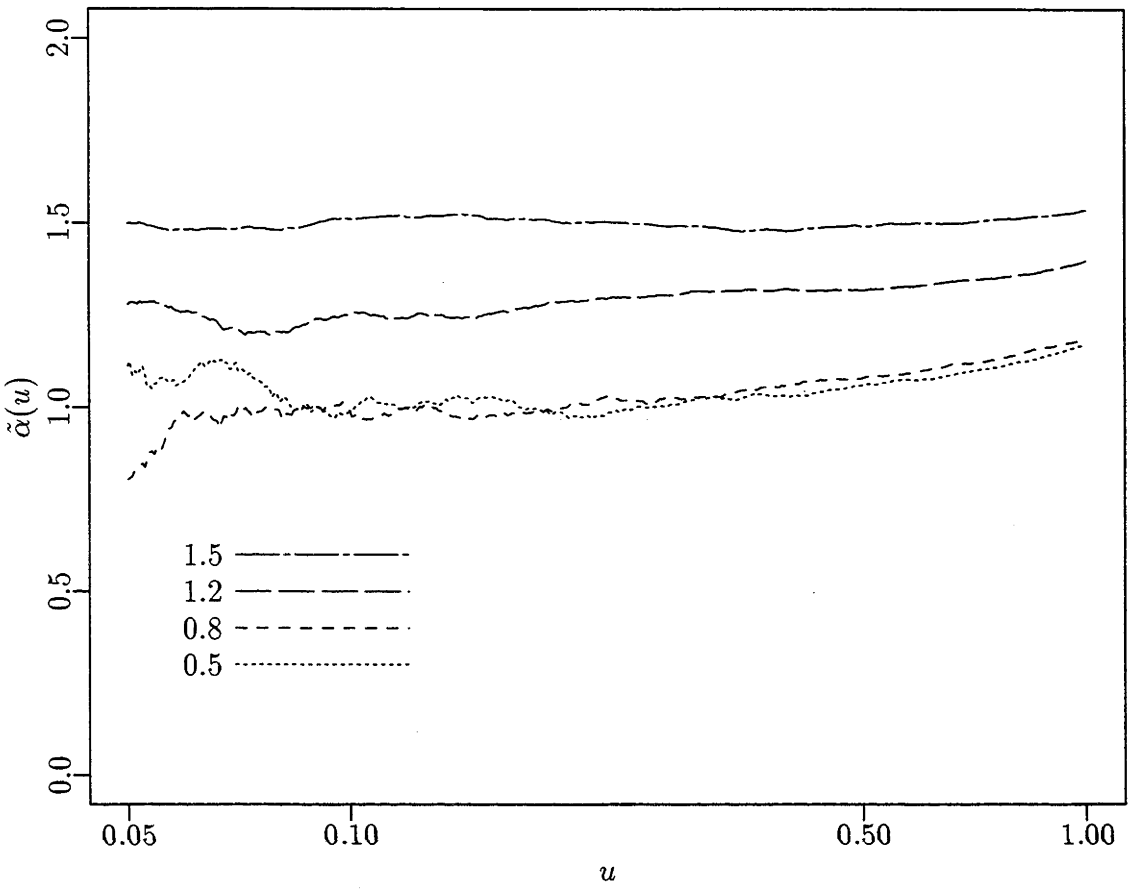


Figure 6.3: Plots of $\tilde{\alpha}(u)$ for the four simulated datasets displayed in Figure 6.2. The scale of u is logarithmic.

6.8.1 Simulated Data without Noise

Figure 6.2 displays typical scatterplots of simulated Poisson-process data in the region $\mathcal{R} = [-1, 1]^2$, with $\alpha = 0.5$, 0.8, 1.2 and 1.5, and the pole at $(0, 0)$. The data were generated by (a) choosing a Poisson-distributed random variable M with mean $200 \times 3^{1/2}$, (b) conditional on M, generating M independent values of the pair (U, V) of independent random variables, where U is Uniform on $[0, 3^{1/2}]$ and V is Uniform on $[0, 1]$, (c) computing the M points obtained by going out a distance $U^{1/(2-\alpha)}$ from the origin and rotating through an angle $2\pi V$ with respect to the x–axis, and (d) retaining only those N points that lay within $\mathcal{R}$. The result is a spherically symmetric Poisson point process with a pole of strength α at the origin

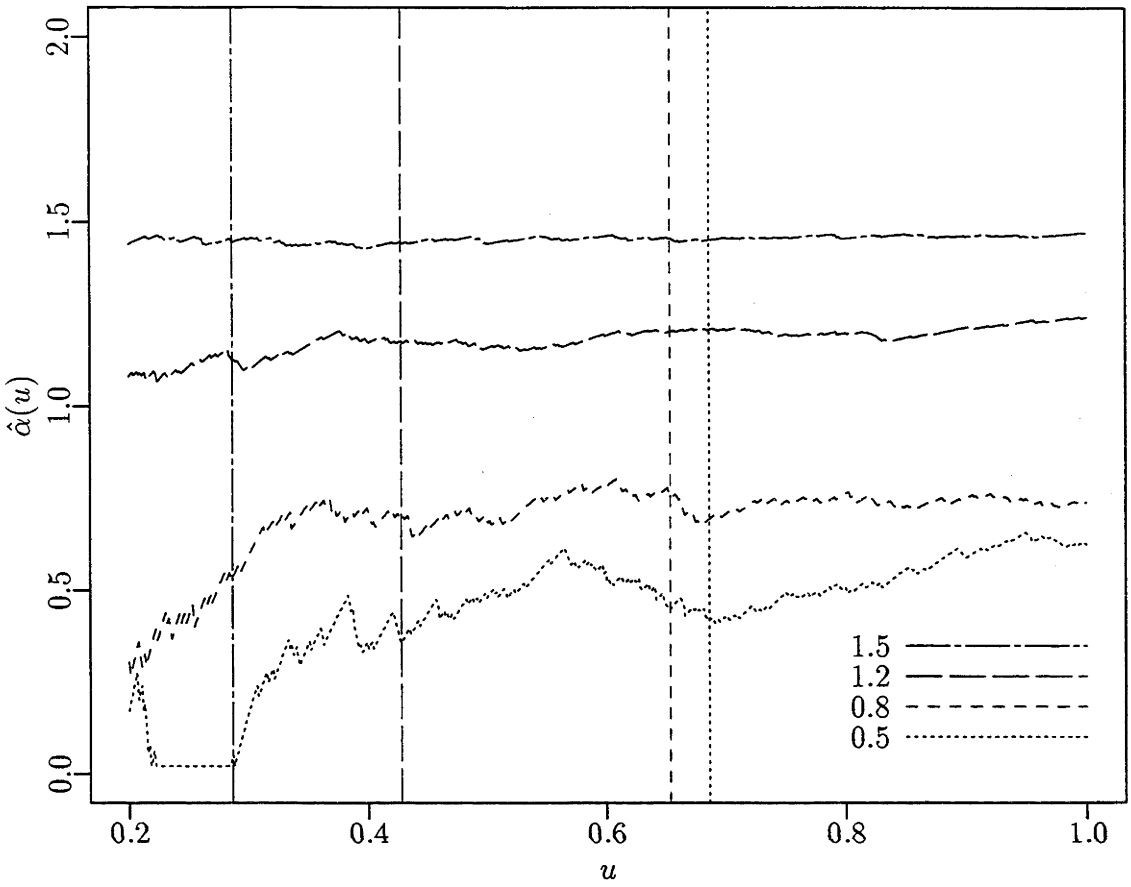


Figure 6.4: Plots of $\hat{\alpha}(u)$ for the four simulated datasets displayed in Figure 6.2. Vertical lines, of the same line type as those for the respective estimates, correspond to that half of the data nearest to $\hat{v}$ being utilised.

and, on average, 200 points within the circle of radius 1 centred at $(0, 0)$. (The Poisson process was generated in a larger region than necessary since, in Section 6.8.2, we shall add noise to it.)

The information depicted in Figures 6.3 and 6.4 is calculated from the data sets presented in Figure 6.2. To estimate the pole, $v = (0, 0)$, we followed a slightly modified version of the prescription in Section 6.4.1. We divided the region $\mathcal{R}$ into a 500×500 grid, and took $\mathcal{S}(w, r)$ to be a closed disc with its centre w on one of the grid points. For each $m = 20, \ldots, 50$, we estimated v as that value of w which allowed the disc to have smallest radius, subject to containing at least m points,

which corresponds to including approximately 10% to 25% of the data in the disc. We took $\hat{v}$ to be the average of these w's over the different values of m.

As expected, the accuracy of the estimates increased with increasing α. The estimated poles for the respective datasets depicted in Figure 6.2 are $(-0.39, 0.010)$, $(0.069, 0.034)$, $(0.027, 0.016)$ and $(-0.0037, -0.0016)$. Bias and mean squared error of estimates of v and α in this section were obtained from a separate simulation study, but for the sake of brevity, we do not report them here.

Figure 6.3 illustrates values of the interpoint distance estimate, $\tilde{\alpha}$, as a function of the threshold parameter, u. As predicted from our theoretical results, the estimator performs poorly when α is close to 1, and breaks down completely when $\alpha < 1$. For α sufficiently greater than 1 the estimator performs well, being surprisingly robust against choice of u. Nevertheless, bias tends to increase with u.

Figure 6.4 shows estimates of $\hat{\alpha}$, defined by minimising the quantity at (6.4), as a function of u, the radius of the larger disc $\mathcal{S}_2$. Vertical lines are drawn through those values of u which correspond to using half the data closest to $\hat{v}$. For simplicity we took the smaller disc $\mathcal{S}_1$ to have zero radius. As indicated by the theory described in Section 6.7, the convergence rate is slower for smaller α. This may be seen from the fluctuations of the estimates for $\alpha = 0.5$ or 0.8.

6.8.2 Simulated Data with Noise

To the Poisson points generated in Section 6.8.1 we added independent Gaussian $\mathrm{N}(0, \sigma I)$ vectors. Figures 6.5 and 6.6 depict the cases $\sigma = 0.01$ and 0.05 respectively. The effects of noise are more prominent for datasets with strong poles, as indicated by the dispersions of data points around the poles. The main effects on the graphs in Figures 6.3 and 6.4 are to pull the curves down and (for Figure 6.4) to increase the slope of the curve. See Figures 6.7 and 6.8, which depict the cases $\tilde{\alpha}(u)$ and $\hat{\alpha}(u)$ respectively. It may be seen from the figures that the curves are pulled much further down for large σ.

To counteract this problem we investigated the method suggested in Section 6.6, which incorporates noise into the approximate likelihood model. Figure 6.9 illustrates the two-parameter log-likelihood surface $\log\{L_1(X_1, \ldots, X_N | \hat{v}, \alpha, \sigma)\}$, with L_1 defined at (6.8), in the case where the true value of σ is 0.05. (We took $\mathcal{S}$ to be the square $[-1, 1]^2$.) Each likelihood surface was obtained by evaluating the log-likelihood function on a 10×10 grid. The integrals in the log-likelihood func-

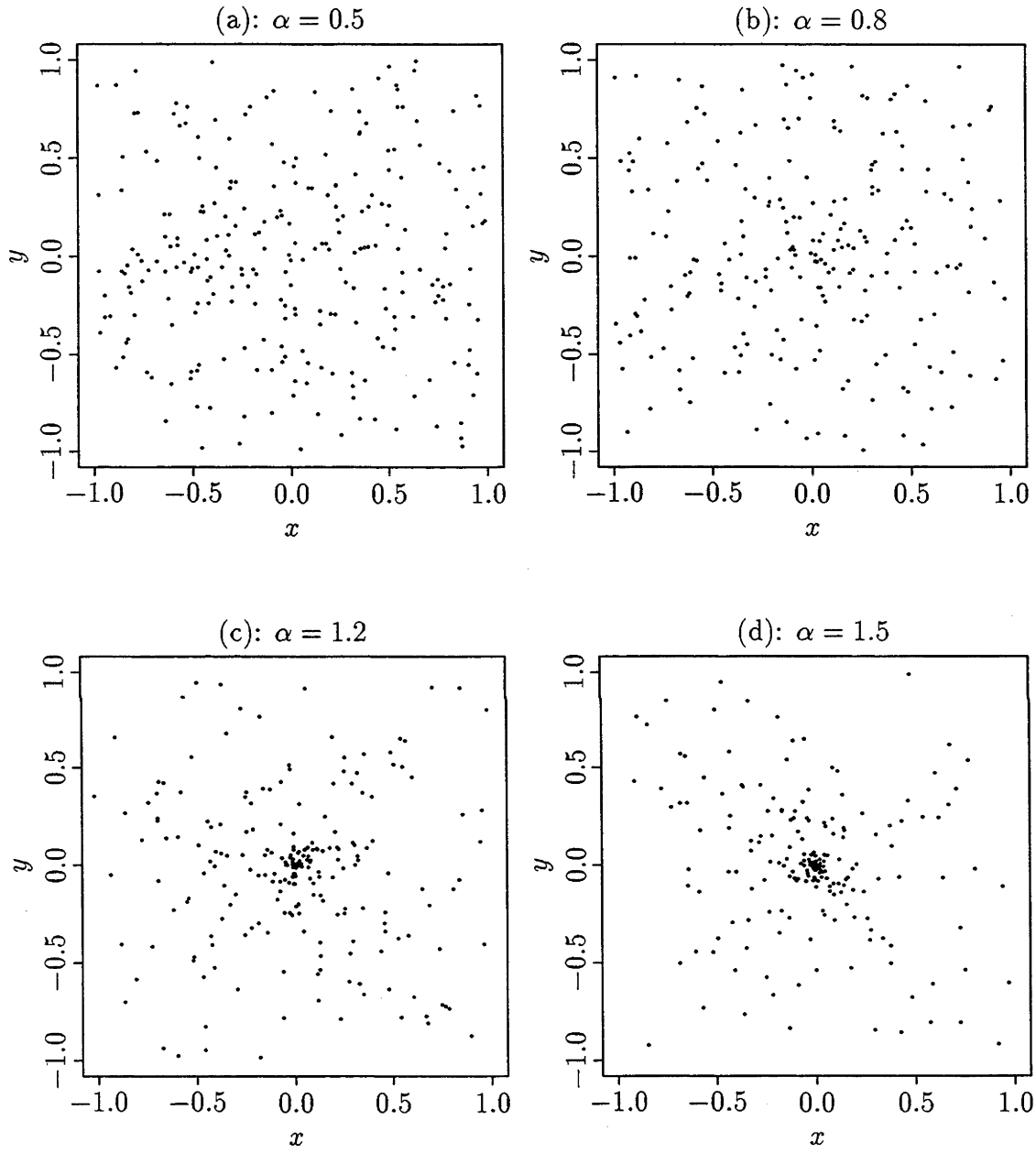


Figure 6.5: Simulated Poisson point process data on $[-1,1]^2$ *with noise* $\sigma = 0.01$. Plots in panels (a)–(d) correspond to pole strength $\alpha = 0.5, 0.8, 1.2$ and 1.5, respectively.

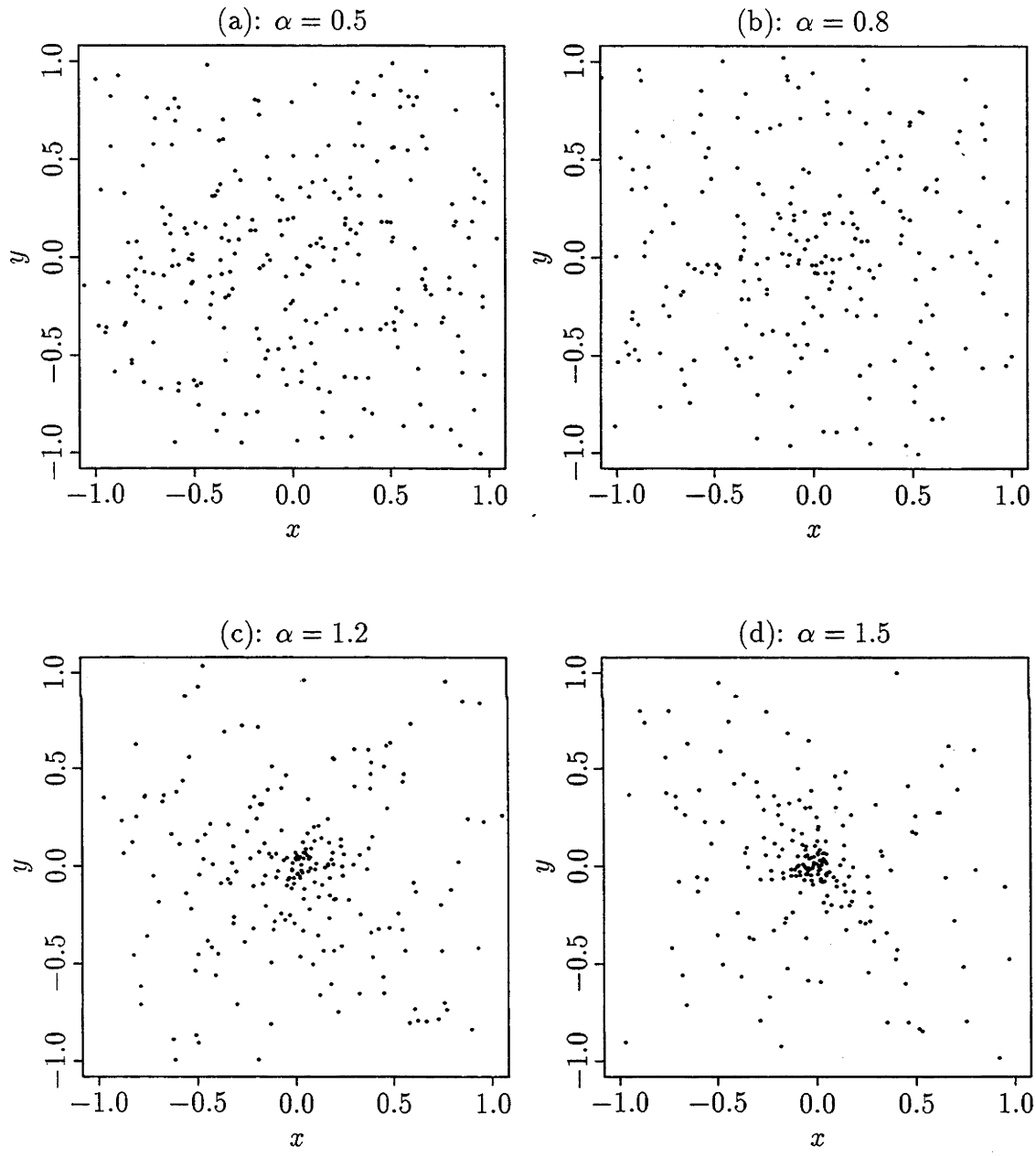


Figure 6.6: Simulated Poisson point process data on $[-1,1]^2$ *with noise* $\sigma = 0.05$. Plots in panels (a)–(d) correspond to pole strength $\alpha = 0.5, 0.8, 1.2$ and 1.5, respectively.

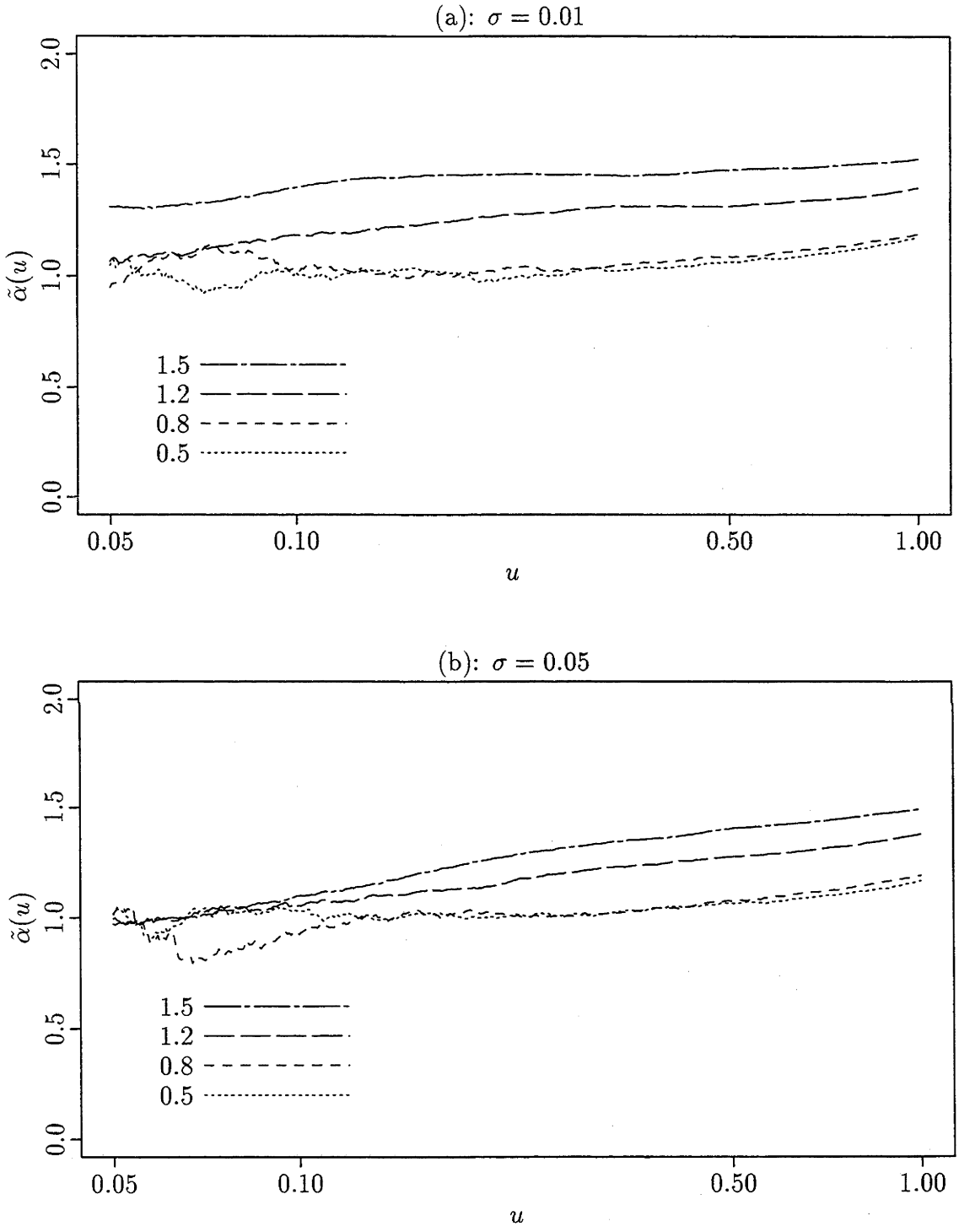


Figure 6.7: Plots of $\tilde{\alpha}(u)$ for noisy Poisson data. Panels (a) and (b) depict the case for $\sigma = 0.01$ and 0.05 respectively.

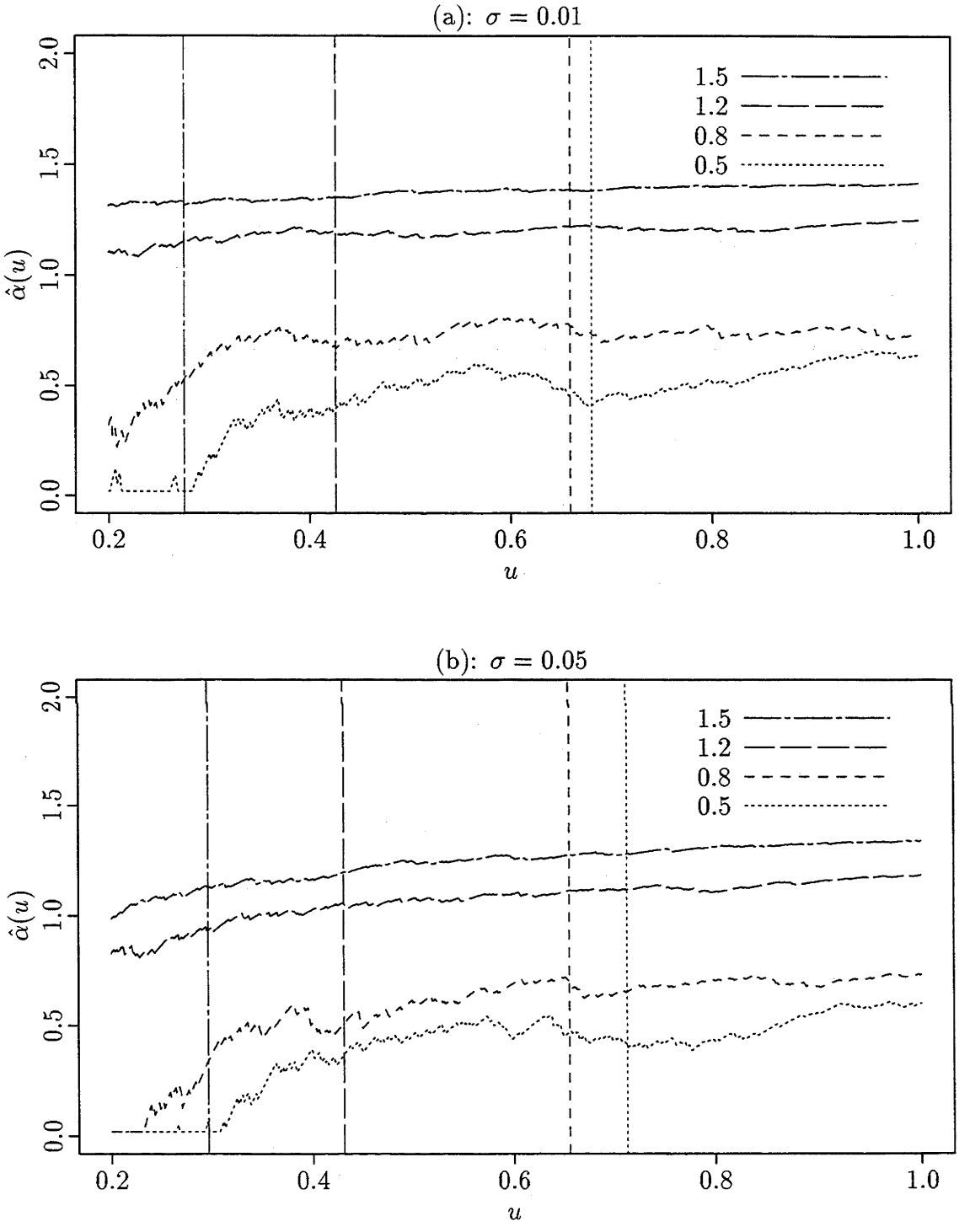


Figure 6.8: Plots of $\hat{\alpha}(u)$ for noisy Poisson data. Panels (a) and (b) depict the case for $\sigma = 0.01$ and 0.05 respectively.

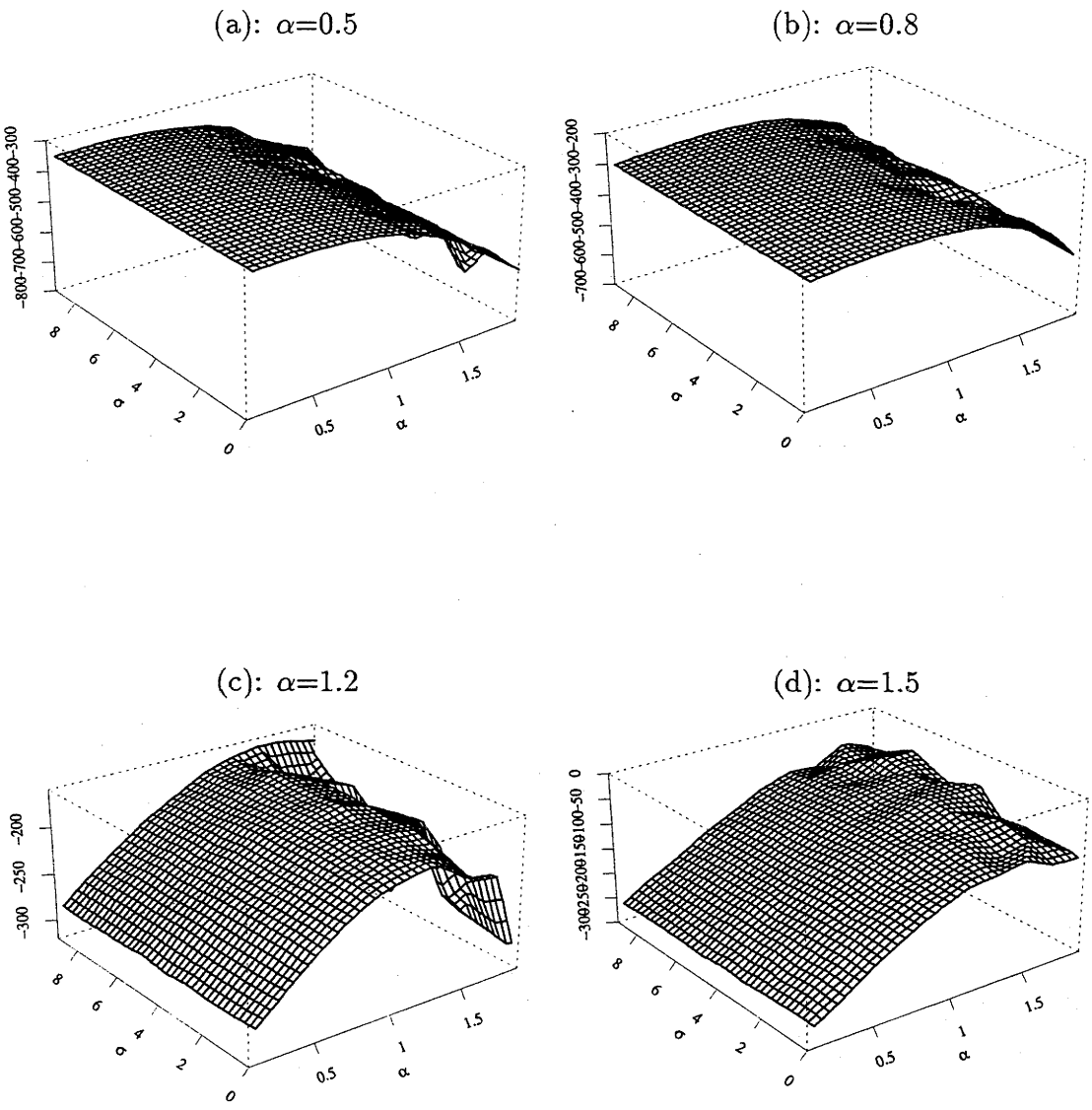


Figure 6.9: Plots of the two-parameter likelihood $L_1(X_1, \ldots, X_N|\hat{v}, \alpha, \sigma)$*, for the data shown in Figure 6.6.* Panels (a)–(d) depict $\alpha = 0.5$, 0.8, 1.2 and 1.5 respectively, in the case $\sigma = 0.05$. The σ-axis is indexed by k, representing k units of $\sigma \times 10^{-2}$.

	(a)			(b)		
α	Bias	Std. Error	MSE	Bias	Std. Error	MSE
0.5	-0.069	0.116	1.8×10^{-2}	-0.049	0.123	1.8×10^{-2}
0.8	-0.055	0.074	8.5×10^{-3}	-0.012	0.076	5.9×10^{-3}
1.2	-0.074	0.063	9.5×10^{-3}	0.012	0.097	9.5×10^{-3}
1.5	-0.059	0.065	7.7×10^{-3}	0.057	0.092	1.2×10^{-2}

Table 6.1: *Bias, standard error and mean squared error of estimates of α, described in Section 6.8.2.* Columns (a) and (b) give estimated bias, standard error and mean squared error of $\hat{\alpha}$, where $\hat{\alpha}$ is obtained by maximising $L_1(X_1, \ldots, X_N | \hat{v}, \alpha, \sigma)$ for $\sigma = 0$ and 0.05 respectively.

tions were computed numerically, using standard quadrature subroutines in the `NAG` library for `Fortran`.

Note particularly the tendency for the likelihood ridge to trend slightly upwards, in the direction of larger α's, as σ increases. That property reflects the rather small amount of information that is available about dispersion in noisy point process data, and indicates that estimation of σ is not a practical proposition. However, given a range of potential values of σ we may compute corresponding estimates of α.

Taking $\hat{\alpha}$ to be the value of α that maximises $L_1(X_1, \ldots, X_N | \hat{v}, \alpha, 0)$ we obtain estimates $\hat{\alpha} = 0.41$, 0.71, 1.23 and 1.46 when the true values are $\alpha = 0.5$, 0.8, 1.2 and 1.5, respectively. If we maximise $L_1(X_1, \ldots, X_N | \hat{v}, \alpha, 0.05)$ we obtain instead $\hat{\alpha} = 0.43$, 0.72, 1.32 and 1.65. Table 6.1 reports estimated bias, standard error and mean squared error of estimates of α.

6.8.3 Kanto Earthquake Data

The spatial distribution of earthquakes in the Kanto region is of great interest to most seismologists because of the complicated nature of plate interactions which occur there, namely the Phillipine Sea, the Eurasian and the Pacific plates. In the region where we analysed pole movements, the situation is further complicated by volcanic activities. The paper by Ogata, Imoto and Katsura (1991) gives more detailed discussion on the tectonic movements in the Kanto District.

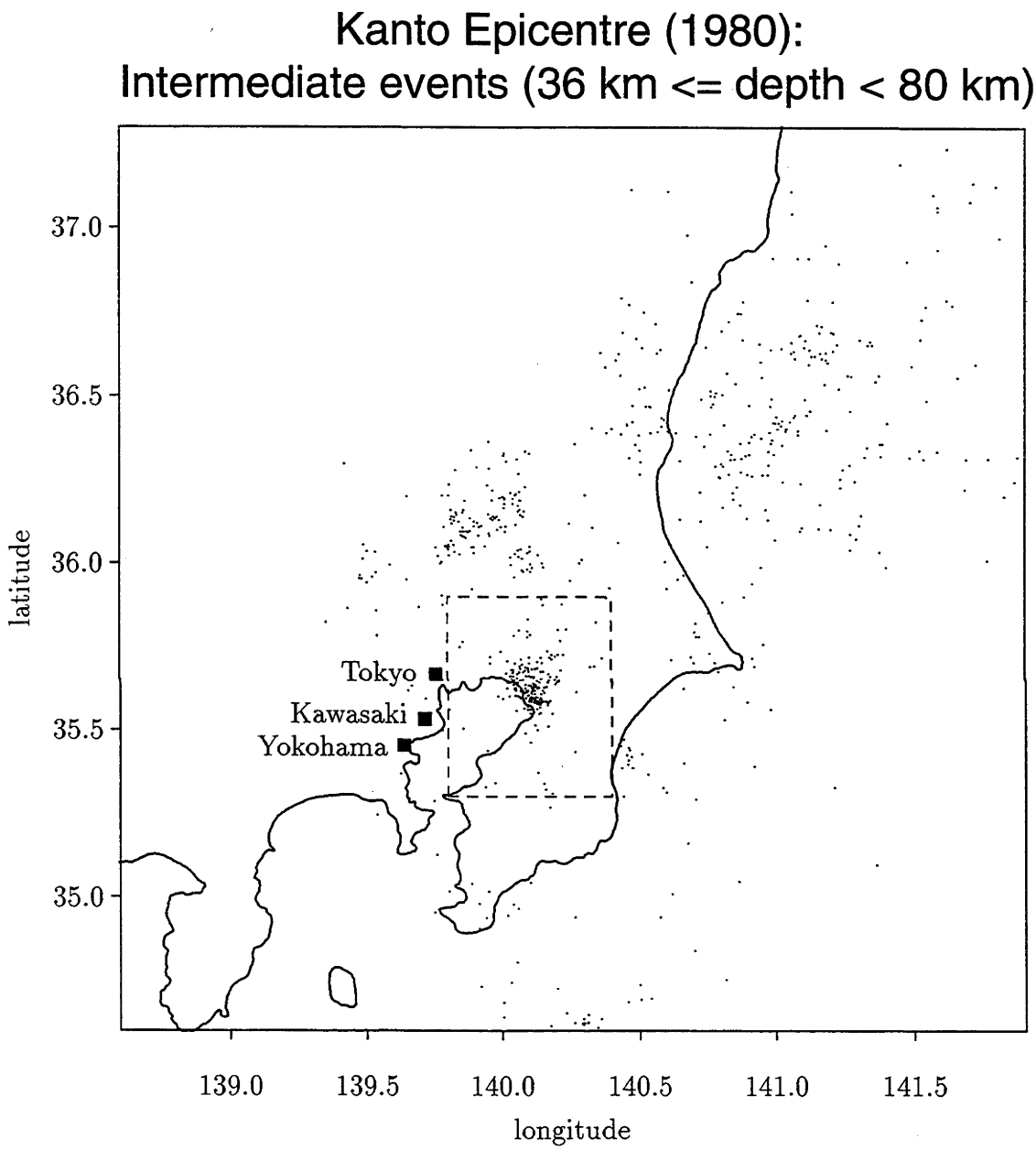


Figure 6.10: Map of 1980 Kanto earthquake epicentres corresponding to magnitude at least 2.2 and depths D_i in the range 36km $\leq D_i <$ 80 km. The box indicates the region chosen for detailed analysis.

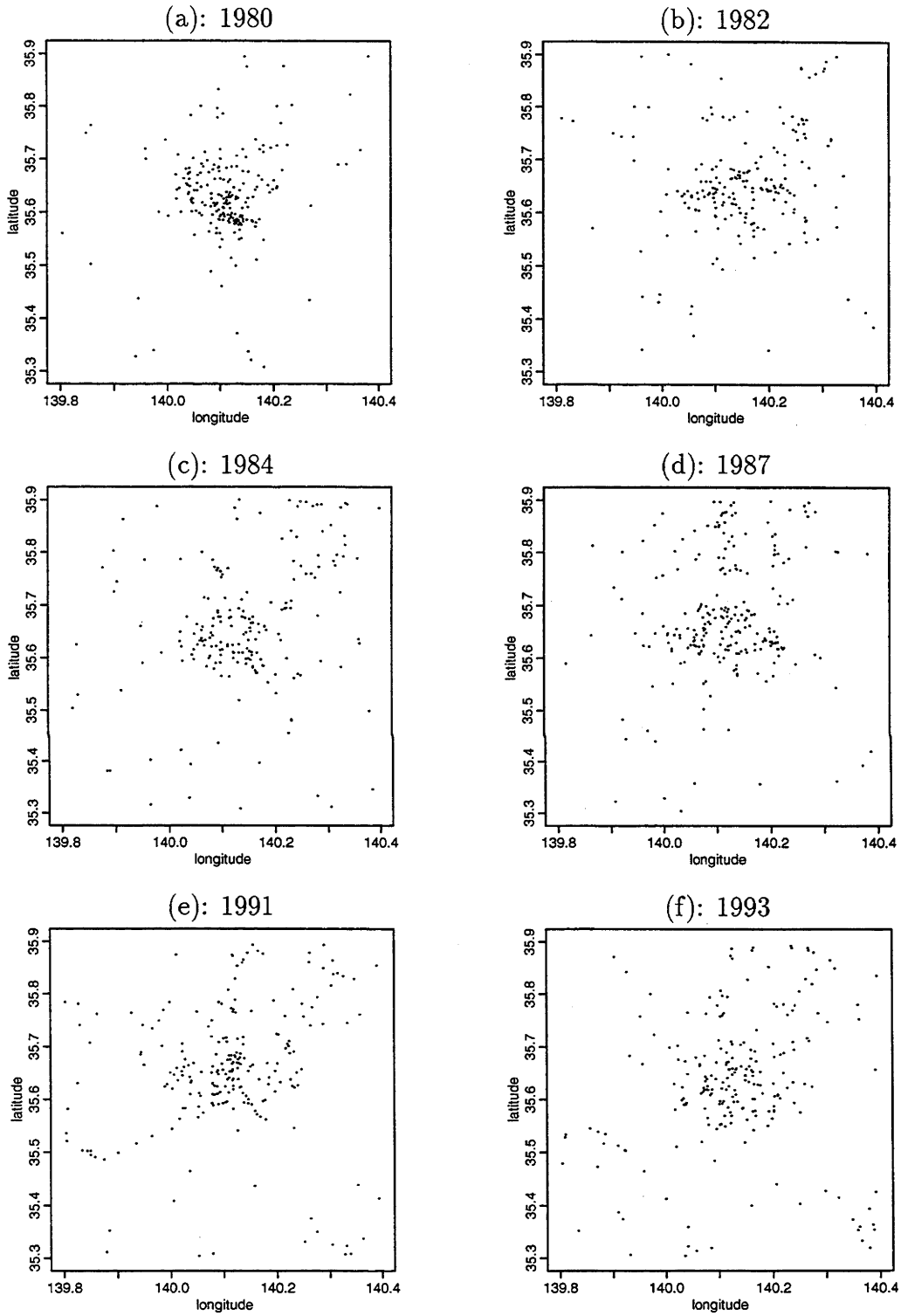


Figure 6.11: Epicentres within the box region for the years 1980, 1982, 1984, 1987, 1991 and 1993. The geographic location of the box region is the same as for Figure 6.10.

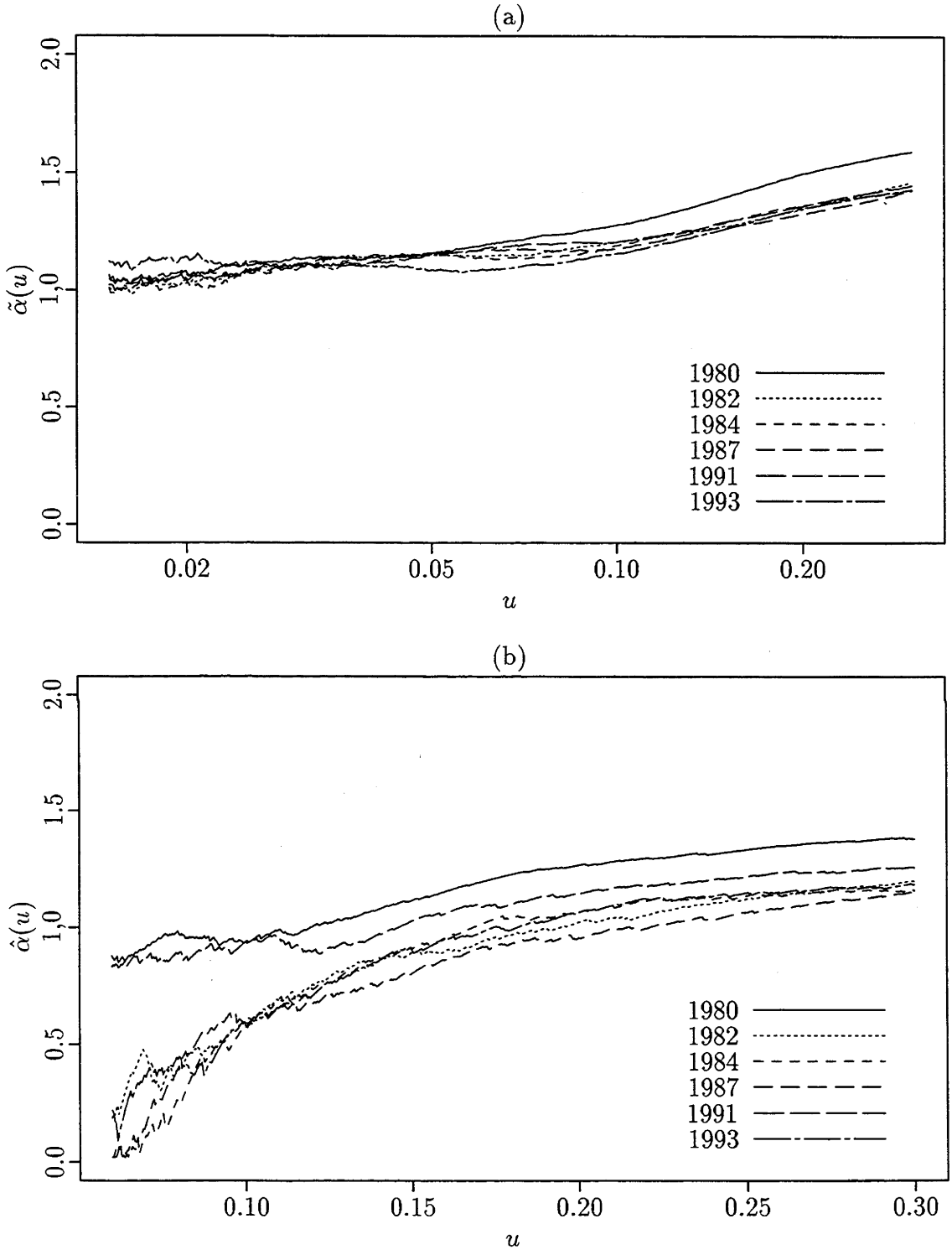


Figure 6.12: Plots of $\tilde{\alpha}(u)$ and $\hat{\alpha}(u)$ for the six years of earthquake data shown in Figure 6.11. Panels (a) and (b) depict $\tilde{\alpha}$ and $\hat{\alpha}$ respectively. We treat the longitudes and latitudes as Cartesian coordinates.

Kanto Intermediate Events (36 km $\leq$ depth $<$ 80 km)

Year	N	min	max	$\hat{v}_x$	$\hat{v}_y$	$\hat{\alpha}$
1980	217	2.20	4.48	140.123	35.596	1.46
				(1.70×10^{-3})	(1.92×10^{-3})	
1982	198	2.20	4.84	140.125	35.645	1.22
				(6.54×10^{-3})	(5.94×10^{-3})	
1984	197	2.00	4.97	140.136	35.632	1.16
				(7.86×10^{-3})	(7.17×10^{-3})	
1987	235	2.00	5.02	140.103	35.660	1.23
				(5.13×10^{-3})	(5.22×10^{-3})	
1991	225	2.00	5.06	140.116	35.654	1.33
				(3.60×10^{-3})	(3.51×10^{-3})	
1993	219	2.00	4.26	140.111	35.629	1.09
				(7.26×10^{-3})	(7.80×10^{-3})	

Table 6.2: *Summary of data and analysis for the Kanto intermediate-depth events.* Minimum and maximum magnitudes, on the Richter scale, are denoted by min and max, respectively. Positions of estimated poles are given by $\hat{v} = (\hat{v}_x, \hat{v}_y)$. The last column lists estimates of α. The bracketted values give estimated standard errors of pole estimates, obtained from simulated data.

We first analysed pole strength, using Kanto earthquake data for the years 1980, 1982, 1984, 1987, 1991 and 1993. A map showing the epicentres for 1980 is displayed in Figure 6.10. We chose to analyse those events which lie between 139.8° and 140.4° longitude and 35.3° and 35.9° latitude, indicated by the box on the map. Tokyo, Kawasaki and Yokohama are situated on the western side of the bay intersected by the box. Data within this box, for all six years, are illustrated in Figure 6.11. Note that because of different scaling of the longitude and latitude axes in Figure 6.10, the boxes in Figure 6.11 are square, whereas those in Figure 6.10 are rectangular. Apart from 1980 and 1982, when only events with magnitude at least 2.2 were included, all events had minimum magnitude 2.0. This ensured that different datasets contained approximately equal numbers of points.

The depths, D_i, of events studied in our analysis were in the range 36 km $\leq D_i <$

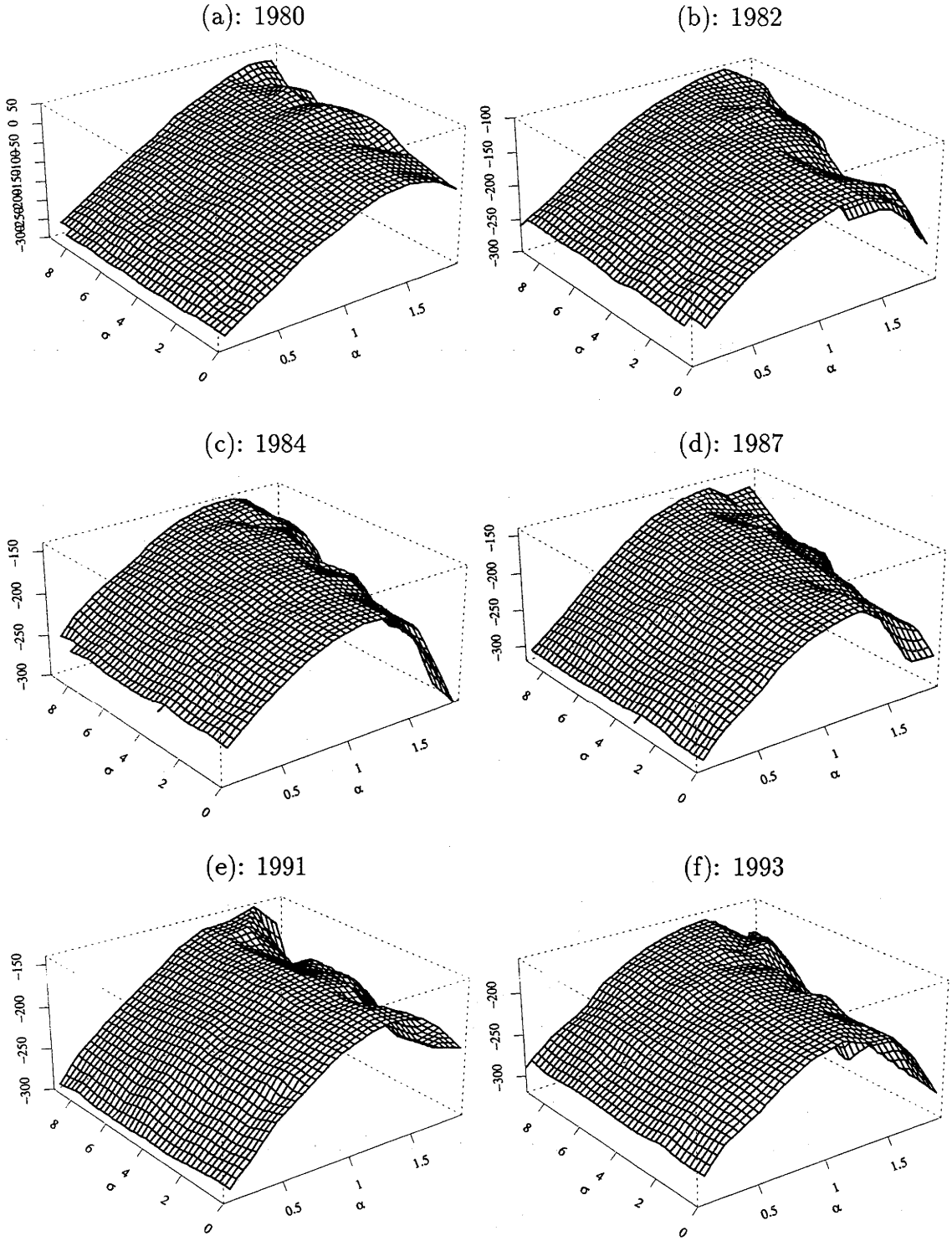


Figure 6.13: Plots of the two-parameter likelihood, L_1, for the Kanto datasets. Results for all six years are shown in panels (a)–(f). The σ-axis is indexed by k, representing k units of σ^{-2}.

Kanto Shallow Events (depth < 36 km)

Year	Longitude	Latitude	N	min	$\hat{v}_x$	$\hat{v}_y$
1980	139.0,139.3	34.8,35.1	222	1.9	139.186	34.966
1983	139.0,139.3	34.8,35.1	238	1.9	139.200	34.938
1984	139.1,139.4	34.8,35.1	383	1.9	139.217	34.928
1986	139.0,139.3	34.8,35.1	207	1.9	139.175	34.948
1987	139.1,139.4	34.8,35.1	489	1.9	139.258	34.913
1988	139.05,139.35	34.8,35.1	237	3.0	139.195	34.951
1989	138.95,139.25	34.8,35.1	175	3.0	139.108	34.986
1993[1]	139.0,139.3	34.8,35.1	337	1.9	139.177	34.937
1993[2]	139.0,139.3	34.8,35.1	614	1.9	139.130	34.976

[1]: includes only events in January
[2]: includes only events in May and June

Table 6.3: *Summary of data and analysis for Kanto shallow events.* The first two columns indicate locations of vertices of the box (i.e. $\mathcal{S}+\hat{v}$) used for detailed analysis. Other notation is as in Table 6.2.

80 km, classified as "intermediate" depths by Harte (1996). Numbers of points, N, in the datasets, and a summary of our results, are given in Table 6.2. More background information about the data may be found in Harte (1996).

Figure 6.12 depicts graphs of $\tilde{\alpha}(u)$ or $\hat{\alpha}(u)$ against u. They show the features that distinguish Figures 6.7 and 6.8 from Figures 6.3 and 6.4, indicating the presence of a small amount of stochastic noise. Therefore, we applied the method proposed in Section 6.6 and trialed in Section 6.8.2. We present the results here for the case where $\mathcal{S} + \hat{v}$ in the denominator of (6.8) is taken to be the box in Figure 6.10. Similar results were obtained with smaller regions, except that there are moderately erratic fluctuations due to small numbers of points in $\mathcal{S}$. Figure 6.13 displays the log-likelihood surfaces, having features broadly similar to those in Figure 6.9. The estimates $\hat{\alpha}$ obtained by maximising $L_1(X_1, \ldots, X_N|\hat{v}, \alpha, 0)$ are given in Table 6.2, as too are the coordinates of $\hat{v} = (\hat{v}_x, \hat{v}_y)$.

Referring to Figure 5.2, we see that by using the smallest interpoint distances, the Hill estimator gives a dimension estimate of around 1.9 for the whole Kanto region.

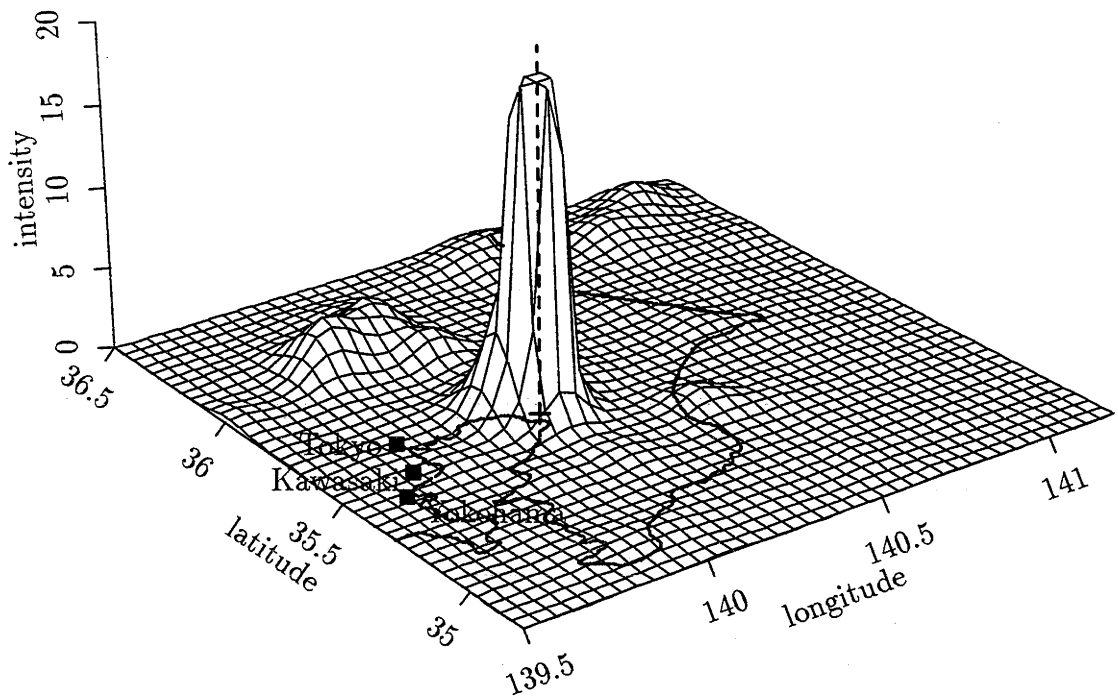


Figure 6.14: Estimate of intensity for data in Figure 6.10 between 139.5° *and* 141.2° *longitude and* 34.8° *and* 36.5° *latitude.* The estimator was constructed using kernel methods, modified to produce the pole. The position of the pole is marked by a cross.

Our interpretation is that the dimension estimate gives a measure of the "average" strength of all different poles in the Kanto region. Using the equation at (6.5), this estimate translates to a pole strength of about 1.05, which may seem relatively low. However, this is consistent with the results given in panel (a) of Figure 6.12, where each curve gives an estimate of pole strength between 1.0 and 1.1. We suspect this is due to the noise effect of the data discussed in the previous paragraph. As larger interpoint distances are employed in measuring the dimension estimate, the interpretation becomes less clear since those distances are mainly contributed by points from different clustered regions. Nevertheless, the gradual decrease in the dimension estimate is consistent with the increase in the estimated pole strength observed in panel (a) of Figure 6.12.

Figure 6.14 depicts a graph of point-process intensity for the 1980 data, illus-

trated in Figure 6.10. Only those events that lie between 139.5° and 141.2° longitude and 34.8° and 36.5° latitude were analysed. The estimator was constructed using kernel methods, with the Gaussian kernel and bandwidth 0.05. Owing to absence of recording stations, data on offshore events can suffer seriously from error. The pole has been artificially created by adding pseudo-data in the vicinity of $\hat{v}$ and reducing the bandwidth to a very small value there.

Similar methods may be applied to shallow Kanto events, corresponding to depths of less than 36 km. Figure 6.15 displays those data, and brief descriptions, together the values of pole estimates, are summarised in Tabele 6.3. Figure 6.16 illustrates the migration of (estimates of) the pole over time. Arrows indicate chronological order, with points representing the years 1980, 1983, 1984, 1986, 1987, 1988, 1989, 1993 (events in January) and 1993 (events in May and June). (No significant poles were apparent in omitted years; and 1993 was a year of unusual seismic activity.) If the data for all these years are plotted together, they appear to be clustered around a short pole line. In this case, however, the line is more appropriately treated as a sequence of simple poles. As pointed out by Professor David Vere-Jones through personal communication, the region in our analysis has significant volcanic activity, and the behaviour observed may not be solely contributed by tectonic movements.

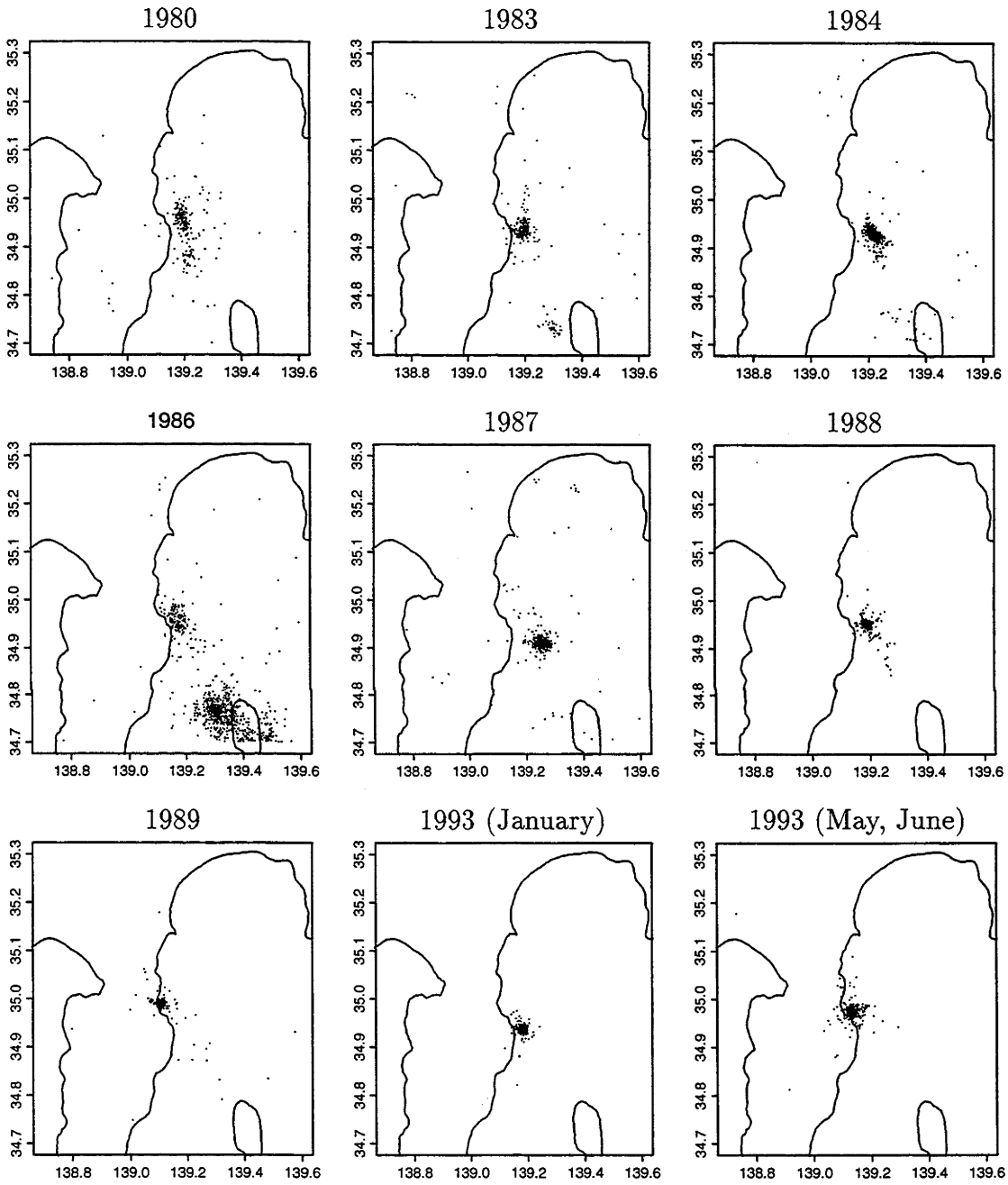


Figure 6.15: Display of Kanto shallow events, 1980–1993. Details of the data are given in Table 6.3.

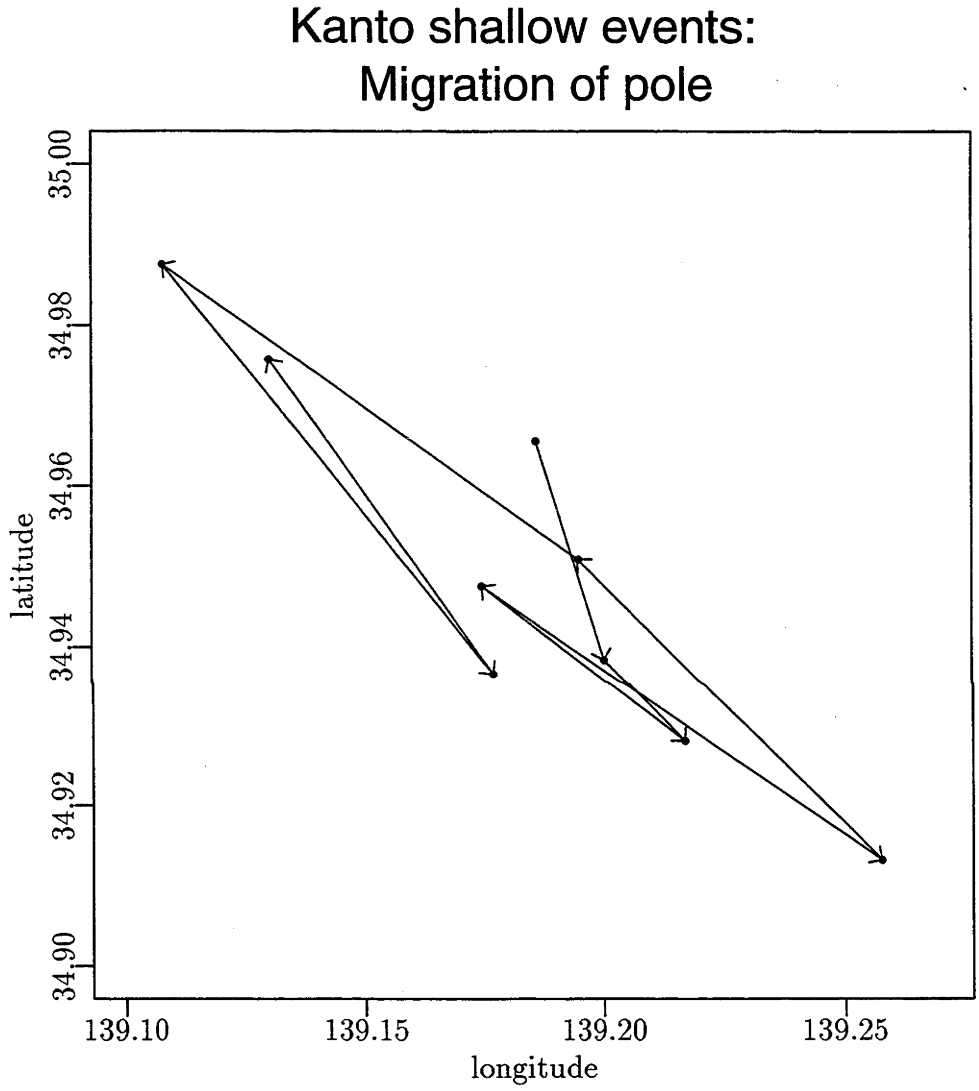


Figure 6.16: Migration of pole for the Kanto shallow events, 1980–1993. The chronological order of the pole occurrences is indicated by arrows. Details of data, which correspond to the years 1980, 1983, 1984, 1985, 1987, 1988, 1989, 1993 (January) and 1993 (May and June), are given in Table 6.3.

References

Abramson, I.S. (1982). On bandwidth variation in kernel estimates – A square root law. *Ann. Statist.* **10**, 1217–1223.

Bartlett, M.S. (1963). Statistical estimation of density functions. *Sankhyā Ser. A* **25**, 245–254.

Breiman, L., Meisel, W. and Purcell, E. (1977). Variable kernel estimates of multivariate densities. *Technometrics* **19**, 135–144.

Cheng, M.Y., Fan, J. and Marron, J.S. (1993). Minimax efficiency of local polynomial fit estimators at boundaries. *Institute of Statistics Mimeo Series #2098*, University of North Carolina at Chapel Hill.

Cheng, M.Y., Hall, P. and Titterington, D.M. (1997). On the shrinkage of local linear curve estimators. *Statistics and Computing* **7**, 11–17.

Chu, C.-K. and Marron, J.S. (1991). Choosing a kernel regression estimator (with discussion). *Statist. Sci.* **6**, 404–436.

Chung, K.L. (1974). *A Course in Probability Theory (2nd ed).* Academic Press, New York.

Clark, R.M. (1977). Nonparametric estimation of a smooth regression function. *J. Roy. Statist. Soc. Ser. B* **39**, 107–113.

Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* **74**, 829–836.

Cleveland, W.S. (1993). *Visualizing Data.* Hobart Press, Summit, N.J.

Cleveland, W.S. and Devlin, S.J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *J. Amer. Statist. Assoc.* **83**, 596–610.

Cleveland, W.S. and Grosse, E.H. (1991). Computational methods for local regres-

sion. *Statist. Comp.* **1**, 47–62.

Cleveland, W.S. and Loader, C.R. (1996). Smoothing by local regression: principles and methods. In *Statistical Theory and Computational Aspects of Smoothing*, W. Härdle and M.G. Schimek (eds.), pp. 10–49. Physica-Verlag, Heidelberg.

Copas, J.B. (1995). Local likelihood based on kernel censoring. *J. Roy. Statist. Soc. Ser. B.* **57**, 221–235.

Cox, D.R. and Isham, V. (1980) *Point Processes.* Chapman and Hall, London.

Cressie, N.A.C. (1993). *Statistics for Spatial Data.* Wiley, New York.

Cutler, C.D. (1991). Some results on the behaviour and estimation of the fractal dimension of distributions on attractors. *J. Statist. Physics* **62**, 651–708.

Cutler, C.D. (1994). A theory of correlation dimension for stationary time series. *Phil. Trans. Roy. Soc. Lond. A* **348**, 343–355.

David, H.A. (1980). *Order Statistics (2nd ed).* Wiley, New York.

Deheuvels, P., Haeusler, E. and Mason, D.M. (1988). Almost sure convergence of the Hill estimator. *Math. Proc. Camb. Phil. Soc.* **104**, 371–381.

Eneva, M. (1996). Effect of limited data sets in evaluating the scaling properties of spatially distributed data: an example from mining induced seismic activity. *Geophys. J. Int.* **124**, 773–786.

Fan, J. (1992). Design-adaptive nonparametric regression. *J. Amer. Statist. Assoc.* **87**, 998–1004.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.* **21**, 196–216.

Fan, J., Farmen, M. and Gijbels, I. (1996). A blueprint of local maximum likelihood estimation. Submitted for publication.

Fan, J. and Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *Ann. Statist.* **20**, 2008–2036.

Fan, J. and Gijbels, I. (1995). Adaptive order polynomial fitting: Bandwidth robustification and bias reduction. *J. Comput. Graphical Statist.* **4**, 213–227.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Application.* Chapman and Hall, London.

Fan, J., Gasser, T., Gijbels., Brockmann, M., and Engel, J. (1997). Local polynomial regression: optimal kernels and asymptotic minimax efficiency. *Ann. Inst. Statist. Math.* **49**, 79–99.

Fan, J., Heckman, N.E. and Wand, M.P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *J. Amer. Statist. Assoc.* **90**, 141–150.

Faraway, J.J. and Jhun, M. (1990). Bootstrap choice of bandwidth for density estimation. *J. Amer. Statist. Assoc.* **85**, 1119–1122.

Gasser, T. and Müller, H.G. (1979). Kernel estimation of regression functions. In *Smoothing techniques for curve estimation,* Lecture Notes in Mathematics 757, pp. 23–69. Springer-Verlag, Berlin.

Gasser, T., Müller, H.G., Köhler, W., Molinari, L., Prader, A. (1984). Nonparametric regression analysis of growth curves. *Ann. Statist.* **12**, 210–229.

Grassberger, P. and Procaccia, I. (1983a). Measuring the strangeness of strange attractors. *Physica D* **9**, 189–208.

Grassberger, P. and Procaccia, I. (1983b). Estimation of the Kolmogorov entropy from a chaotic signal. *Phys. Rev. A* **28**, 2591–2593.

Grassberger, P. and Procaccia, I. (1983c). Characterisation of strange attractors. *Phys. Rev. Letters* **50**, 346–349.

Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach.* Chapman and Hall, London.

Härdle, W. (1986). A note on jackknifing kernel regression function estimators. *IEEE Trans. Inf. Theory* **32**, 298–300.

Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge University Press, Cambridge, U.K.

Härdle, W., Hall, P. and Marron, J.S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? (with discussion). *J. Amer. Statist. Assoc.* **83**, 86–101.

Härdle, W., Hall, P. and Marron, J.S. (1992). Regression smoothing parameters that are not far from their optimum. *J. Amer. Statist. Assoc.* **87**, 227–233.

Härdle, W. and Marron, J.S. (1995). Fast and simple scatterplot smoothing. *Com-*

put. Statist. and Data Analysis **20**, 1–17.

Hall, P. (1978). Representations and limit theorems for extreme value distributions. *J. Appl. Prob.* **15**, 639–644.

Hall, P. (1982). On some simple estimates of an exponent of regular variation. *J. Roy. Statist. Soc. Ser. B* **44**, 37–42.

Hall, P. (1990). On the bias of variable bandwidth curve estimators. *Biometrika* **77**, 529–535.

Hall, P. and Marron, J.S. (1997). On the role of the ridge parameter in local linear smoothing. *Probab. Theory Related Fields* **108**, 495–516.

Hall, P. and Turlach, B.A. (1997a). Enhancing convolution and interpolation methods for nonparametric regression. *Biometrika* **84**, 779–790.

Hall, P. and Turlach, B.A. (1997b). Interpolation methods for adapting to sparse design in nonparametric regression (with discussion). *J. Amer. Statist. Assoc.* **92**, 466–476.

Hall, P. and Wehrly, T.E. (1991). A geometrical method for removing edge effects from kernel-type nonparametric regression estimators. *J. Amer. Statist. Assoc.* **86**, 665–672.

Hart, J.D. (1997). *Nonparametric smoothing and lack-of-fit tests.* Springer-Verlag, New York.

Harte, D.S. (1996). *Multifractals — Theory and Applications.* PhD thesis, Victoria University of Wellington.

Hastie, T.J. and Loader, C.R. (1993). Local regression: automatic kernel carpentry (with discussion). *Statist. Sci.* **8**, 120–143.

Hentschel, H.G.E. and Procaccia, I. (1983). The infinite number of generalized dimensions of fractals and strange attractors. *Physica D* **8**, 435–444.

Herrmann, E. (1996). On the convolution type kernel regression estimator. *Preprint No. 1833,* Fachbereich Mathematik, TH Darmstadt, Germany.

Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution. *Ann. Statist.* **3**, 1163–1174.

Hirata, T. and Imoto, M. (1991). Multifractal analysis of spatial distribution of microearthquakes in the Kanto region. *Geophys. J. Int.* **107**, 155–162.

Hjort, H.L. (1991). Semiparametric estimation of parametric hazard rates. In *Survival Analysis: State of the Arts,* P.S. Goel and J.P. Klein (eds.), pp. 211–236. Kluwer, Dordrecht.

Hjort, H.L. (1997). Dynamic likelihood hazard rate estimation. *Biometrika* **84**, to appear.

Hjort, H.L. and Glad, I.K. (1995). Nonparametric density estimation with a parametric start. *Ann. Statist.* **23**, 882–904.

Hjort, H.L. and Jones, M.C. (1996). Locally parametric nonparametric density estimation. *Ann. Statist.* **24**, 1619–1647.

Jones, M.C. (1993). Simple boundary correction for kernel density estimation. *Statist. Comp.* **3**, 135–146.

Jones, M.C. and Foster, P.J. (1993). Generalized jackknifing and higher order kernels. *J. Nonpara. Statist.* **3**, 81–94.

Jones, M.C., Linton, O. and Nielsen, J.P. (1995). A simple bias reduction method for density estimation. *Biometrika* **82**, 327–338.

Jones, M.C. and Signorini, D.F. (1997). A comparison of higher-order bias kernel density estimators. *J. Amer. Statist. Assoc.* **92**, 1063–1073.

Kingman, J.F.C. (1993). *Poisson Processes.* Oxford University Press, Oxford, U.K.

Lejeune, M. and Sarda, P. (1992). Smooth estimators of distribution and density functions. *Comput. Statist. and Data Analysis* **14**, 457–471.

Linton, O. and Nielsen, J.P. (1994). A multiplicative bias reduction method for nonparametric regression. *Statist. Prob. Letters* **19**, 181–187.

Loader, C.R. (1996). Local likelihood density estimation. *Ann. Statist.* **24**, 1602–1618.

Macauley, F.R. (1931). *The Smoothing of Time Series.* National Bureau of Economic Research, New York.

Marron, J.S. and Wand, M.P. (1992). Exact mean integrated squared error. *Ann. Statist.* **20**, 712–736.

Mason, D.M. (1982). Laws of large numbers for sums of extreme values. *Ann. Prob.* **10**, 754–764.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models (2nd ed).* Chap-

man and Hall, London.

Mikosch, T. and Wang, Q. (1993). Some results on estimating Rényi type dimensions. Institute of Statistics and Operations Research Report, Victoria University of Wellington.

Mikosch, T. and Wang, Q. (1995). A Monte-Carlo method for estimating the correlation exponent. *J. Statist. Physics* **78**, 799–813.

Miller, R.G. (1981). *Survival Analysis.* Wiley, New York.

Müller, H.G. (1984). Boundary effects in nonparametric curve estimation models. In *COMPSTAT*, pp. 84–89. Physica-Verlag, Heidelberg.

Müller, H.G. (1991). Smooth optimum kernel estimators near endpoints. *Biometrika* **78**, 521–530.

Nadaraya, E.A. (1964). On estimating regression. *Theory Probab. Applic.* **15**, 134–137.

Nason, G.P. and Silverman, B.W. (1997). Wavelets for regression and other statistical problems. In *Smoothing and Regression: Approaches, Computation and Application*, M.G. Schimek (ed.), Wiley. To appear.

Ogata, Y., Imoto, M. and Katsura, K. (1991). 3-D spatial variation of *b*-values of magnitude-frequency distribution beneath the Kanto District, Japan. *Geophys. J. Int.* **104**, 135–146.

Olkin, I. and Spiegelman, C.H. (1987). A semiparametric approach to density estimation. *J. Amer. Statist. Assoc.* **82**, 858–865.

Parzen, E. (1962). On the estimation of a probability density function and the mode. *Ann. Math. Statist.* **33**, 1065–1076.

Pesin, Y.B. (1993). On rigorous mathematical definitions of correlation dimension and generalized spectrum for dimensions. *J. Statist. Physics* **71**, 529–547.

Pollard, D. (1984). *Convergence of Stochastic Processes.* Springer-Verlag, New York.

Ramsay, J.O. and Silverman, B.W. (1997). *Functional Data Analysis.* Springer-Verlag, New York.

Reiss, R.-D. (1989). *Approximate Distributions of Order Statistics: With Applications to Nonparametric Statistics.* Springer-Verlag, New York.

Rice, J. (1984). Boundary modification for kernel regression. *Commun. Statist.* **13**, 893–900.

Ripley, B.D. (1981). *Spatial Statistics.* Wiley, New York.

Ruppert, D. and Wand, M.P. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.* **22**, 1346–1370.

Schucany, W.R. and Sommers, J.P. (1977). Improvement of kernel type density estimators. *J. Amer. Statist. Assoc.* **72**, 420–423.

Seifert, B. and Gasser, T. (1996a). Finite sample variance of local polynomials: Analysis and solutions. *J. Amer. Statist. Assoc.* **91**, 267–275.

Seifert, B. and Gasser, T. (1996b). Variance properties of local polynomials and ensuing modifications. In *Statistical Theory and Computational Aspects of Smoothing*, W. Härdle and M.G. Schimek (eds.), pp. 50–79. Physica-Verlag, Heidelberg.

Shao, J. (1991). Second-order differentiability and jackknife. *Statist. Sinica* **1**, 185–202.

Shorack, G.R. and Wellner, J.A. (1986). *Empirical processes with applications to statistics.* Wiley, New York.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

Smith, R.L. (1992). Estimating dimension in noisy chaotic time series. *J. Roy. Statist. Soc. Ser. B* **54**, 329–351.

Stone, C.J. (1977). Consistent nonparametric regression (with discussion). *Ann. Statist.* **5**, 595–645.

Takens, F. (1985). On the numerical determination on the dimension of an attractor. In *Dynamical Systems and Bifurcations. Lecture Notes in Mathematics, 1125.* B.L.J. Braaksma, H.W. Broer and F. Takens (eds.), pp. 99–106. Springer-Verlag, Berlin.

Theiler, J. (1988). Lacunarity in a best estimator of fractal dimension. *Physics Letters A* **133**, 195–200.

Theiler, J. (1990). Statistical precision of dimension estimators. *Phys. Rev. A* **41**, 3038–3051.

Tsybakov, A.B. (1986). Robust function reconstruction by local approximation. *Problems of Information Transmission* **22**, 69–84.

Uspensky, J.V. (1937). *Introduction to Mathematical Probability.* McGraw-Hill, New York.

Vere-Jones, D. (1996). On the fractal dimensions of point process. Institute of Statistics and Operations Research, Victoria University of Wellington.

Vere-Jones, D., Davies, R.B., Harte, D., Mikosch, T. and Wang, Q. (1997). Problems and examples in the estimation of fractal dimension from meteorological and earthquake data. In *Applications of Time Series Analysis in Astronomy and Meteorology,* T. Subba Rao, M.B. Priestley and O. Lessi (eds.), pp. 359–375. Chapman and Hall, London.

Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing.* Chapman and Hall, London.

Watson, G.S. (1964). Smooth regression analysis. *Sankhyā Ser. A* **26**, 359–372.